Google Search in India: Unveiling the Geo-Personalized Web

Pranav Chatur*
pranavchatur@gmail.com
The LNM Institute of Information Technology
Jaipur, Rajasthan, India

Naman Jain*
jainnj2210@gmail.com
The LNM Institute of Information Technology
Jaipur, Rajasthan, India

Varun Trivedi*
varun.trivedi1803@gmail.com
The LNM Institute of Information Technology
Jaipur, Rajasthan, India

Ayush Dhoot*
ayush01.dhoot@gmail.com
The LNM Institute of Information Technology
Jaipur, Rajasthan, India

Sakthi Balan Muthiah
sakthi.balan@lnmiit.ac.in
The LNM Institute of Information Technology
Jaipur, Rajasthan, India

ABSTRACT

The ubiquitous presence of search engines has revolutionized the way people access information. Google, as the dominant search engine worldwide, plays a pivotal role in shaping information retrieval experiences. It employs personalized search algorithms to deliver tailored search results based on each user's preferences. Despite numerous studies on general personalization in search engines, there is limited research on geolocation-driven personalization in search engine results, particularly in India. This research paper aims to quantitatively analyze and assess the impact of geolocation on personalized search results within the context of India.

To conduct this study, we have selected an extensive set of search queries across various domains. Multiple geolocations within India were chosen to represent different regions, cities, and rural areas. Using a systemic methodology, we collected and analyzed search results for each query, keeping the user's geolocation as a variable. The study focuses on the extent of personalization introduced by Google's search algorithms in search result rankings based on geolocation. The findings indicate that personalization influences search results, though the degree of variation depends on the specific search query category and result ranking. Queries regarding popular or local items show higher personalization, while within-state personalization is more elevated in larger states or cities with cosmopolitan populations.

This research paves the way for fostering a deeper understanding of the implications of geolocation-driven search result personalization.

CCS CONCEPTS

• **Social and professional topics** → **Geographic characteristics**; Cultural characteristics; • **General and reference** → *Empirical studies*; • **Information systems** → **Similarity measures**.

*All authors contributed equally to this research.

CODS COMAD, January 2024, Bangalore, Karnataka, India

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

KEYWORDS

Geolocation, Rank-Biased Overlap(RBO), Google search engine, Web search personalization, Internet Filter Bubble

1 INTRODUCTION

The Internet has a rich history spanning several decades, but in the last decade its availability has expanded rapidly, particularly in developing countries. India has seen a rapid increase in internet users in this decade. A report [11] published by IAMAI and Kantar in 2022, revealed that a majority of Indians have become active internet users, with 52% of the population or 759 million people accessing the internet at least once a month. This widespread accessibility has revolutionized how people live and interact with the world.

At the heart of this dependency lie Search Engines, acting as the gateway to the vast realm of information and services the internet offers. Search engines have become an integral part of a person's daily routine, aiding in research, shopping, communication, entertainment, and much more. They have streamlined how we access and consume information, becoming an indispensable tool for navigating the digital landscape. It is no surprise that Google has been the most visited search engine on the Internet for several years now; the latest data shows that Google processes over 99,000 searches every single second, which makes more than 8.5 billion searches a day [15].

Major search engines like Google use personalization, which allows them to personalize the results or their order depending on the user who is submitting the query, to better cater to their needs. With so much reliance on the Google search engine, the algorithmic bias of the search results provided by Google must be analyzed, as any discrepancy can mislead the user, causing much bigger problems. Personalization can be helpful to a certain extent, but it must not distort the data and misrepresent information or keep some information hidden from the user stating that it is irrelevant. (Filter Bubble Effect) [16]. In the Indian context, apart from scattered attempts, there has been a lack of academic research on the topic of personalization in the search results provided by Google.

Motivated by the above concerns, we study how geolocation affects personalization in Google Search within India. The first of its kind in India, this study aims to provide in-depth and

unprecedented information on the geolocation-based personalization of Google search results.

For this, we start by putting different query terms in the Google search engine from various geographical locations and then scraping the hyperlinks of the results. We automated this process using Python and one of its frameworks, Selenium. For our target geolocations, we divided our research into two granularities, one of which focused on comparing each of the Indian state capitals with the National capital, while the other focused on collecting the data for about 4-5 cities of a state and comparing it with the state's capital.

Next, we focused on assembling an extensive set of search terms (or queries) and the geographical locations for our research to calculate the personalization. We created a list of 54 search terms under various categories that cover aspects such as 'Medical', 'Controversial', 'Employment', 'Education', 'Expected Local', 'Notable Entities', 'Sports' & 'Current Affairs'. We put these queries over a collection of target geolocation based all over the Indian demographic in the Google search engine.

Then, we analyzed the collected data. The approach followed was a black box approach. The study focused on the input and the acquired output. In this case, the queries and Google search results. We have used RBO (Ranked Biased Overlap) [29] index to get a quantitative analysis of the level of personalization for a location. For visualizing the results, we have used standard plots like box plots and bar graphs as well as a unique method of representation where we show the personalization in a query using color gradients on the map of India. The box plot helps us to represent the collected data for each city, and aggregate the RBO values for all queries tested distinctly.

The paper is structured as follows: Section 2 examines relevant literature and research, Section 3 outlines our data collection methodology, and Sections 4 and 5 present our analysis and findings. Future research directions and open challenges are discussed in Section 6, and the paper concludes in Section 7.

2 LITERATURE REVIEW

In this section, we have examined and analyzed various literature and research papers related to the personalization of web searches and the factors that affect it. This topic has been studied by several researchers around the globe [14, 17, 19, 21, 23–26].

Personalized search is web search results that are tailored specifically to a user's interests by incorporating information about the individual beyond the specific query provided.

In their research, Dou et al. [5], examined different ways to personalize search results, evaluating five personalized search strategies, including two click-based, and three profile-based strategies. The paper suggests that profile-based personalized search strategies are less reliable than click-based ones. On certain searches, they may increase search accuracy, but on many others, they could undermine it.

However, there is very little concrete information about how the biggest search engines such as Google personalize their search results.

Several studies have looked into various parameters that can affect personalization, such as the impact of geolocation on personalization [1, 2, 30] and the use of browser histories to estimate user demographics which may then be used to personalize content [6, 10].

Hannak et al. [13] in their work wrote that search engines are the primary gateway to information in the developed world and personalization of search results has led to worries about the Filter Bubble effect, where the algorithm decides that some useful information is irrelevant to the user, and thus prevents them from locating it [16]. Motivated by concerns about Filter Bubble Effect, their prior work [9] set out to explore various factors which triggered personalization in Google Search. The factors, they explored were Basic Cookie Tracking, Browser User-Agent, IP Address Geolocation, Inferred Geolocation, Browsing History, Search-Result-Click History, Search History, and User Profile Attributes. They found that Google infers users' geolocation based on their IP address and that location-based personalization caused more differences in search results than any other single feature.

In [13], the authors collected search results for 30 days from Google Search in response to 240 different queries. By comparing search results gathered from 59 GPS coordinates around the US at three different levels (county, state, and national). They observed that differences in search results due to personalization grow as physical distance increases. However, these differences are dependent on what a user searches for queries for local establishments receive 4-5 different results per page, while more general terms exhibit essentially no personalization.

Given growing concerns about the Filter Bubble effects and Search Engine Optimization [28], this area seems promising for future research.

Our research focuses on quantifying the amount of personalization induced by Google's search engine due to changes in geolocations in India, i.e. when the same query is searched from different geographical locations. We have also composed a comprehensive set of queries(search terms) to find out what type of queries are most affected by the change in geolocation.

3 METHODOLOGY

3.1 Locations

To measure the extent of personalization found in the results of search terms across different geolocations, we have gathered data over two levels, the National level and State level, to capture disparate results.

The first set of locations, curated for a National level study, includes the capital cities of all the states of India and the capital city of India, New Delhi. This choice of locations is based upon the assumption that the capital city for a particular state precisely summarises the different notions embodied by all the citizens of the respective Indian state[1]. The second set of locations, (mentioned in Table 1), chosen to challenge the above assumption, constitutes a state-wide study, where for every state of India, a set of 4-5 locations within the state is chosen. This includes the capital of the state as a Tier-1 city and the remaining cities are selected

[1]Author's note: At the time of the study, the country of India is divided into 28 states and 8 union territories but considering the geographical size of Ladakh and Jammu & Kashmir, the 2 union territories have been included as states to observe the personalization imparted by the Google search engine on citizens of those union territories.

based on the population density of each city within their respective states [8]. We have carefully chosen two cities with the lowest population density and accessible geo-targets, while the remaining two cities are selected based on their high population density [18]. The purpose of choosing such a set of locations is to measure the intra-state personalization observed across the different states of India and mathematically realize the difference in web results observed by people living in urban and rural environments for every Indian state.

3.2 Search Terms

A search term, or a query, is composed of one or more keywords that are to be searched across the different geolocations. We have come up with a list of search terms (some of them mentioned in Table 2) spread across multi-disciplinary fields to be searched across the different geolocations mentioned in the previous sections.

Having an eclectic set of search terms is critical for capturing the personalization observed across a wide range of topics like Employment, Medical, Current Affairs, and many such. Therefore the queries have been tagged with words such that they can be categorized. Tagging is a better option as compared to mutually exclusive categorization as search terms can often have multiple domains of interest and tagging helps consider them across all domains of interest. Virtually it can be considered the same as categorization; however, each query can have more than one tag, while a query can have just one category. Therefore, tagging is our preferred choice of aggregation. The primary purpose is aggregation so that personalization can be observed across a wide range of topics without looking into any particular search term in an isolated manner.

3.3 Implementation and Algorithm

3.3.1 *Scraping the data from Google Webpages*

. At its core, our experiment revolves around gathering data for each location corresponding to every search term. The data collection for the National-level (Table 1) and State-level (Table 3) study took place on the 19th of January 2023 and the 20th of March 2023 respectively. Delineating the exact form of data we have collected, we have systematically extracted all web links featured on the initial page of Google search results (Refer Figure 1).

With the physical relocation of hardware for every geolocation being infeasible, the task here is to collect the set of links for all the different geolocations. To achieve this, we have implemented an innovative technique.

An illustrative URL for a Google search takes the form as follows: *https://www.google.co.in/search?q=example+query&uule=w+examplecity*

For our research objective, we have leveraged two primary query parameters embedded within Google's search URLs, namely, '*q*' and '*uule*' [22]. Parameter '*q*' is used to search for a particular search term and parameter '*uule*' can be used for changing to the required geolocation. Here the search term is 'example query' and geolocation is 'examplecity'. We have used Google Ads API Geo targets [7] to get the exact name for every intended geo target as an argument to parameter '*uule*' (Refer Figure 2).

Table 1: Geolocations for quantifying bias observed within the different states of India

State	Chosen cities
Andhra Pradesh	Tadipatri, Kadiri, Guntur, Vijayawada, Visakhapatnam
*Jammu and Kashmir	
*Ladakh, India	
Punjab	Sunam, Barnala, Faridkot, Jalandhar, Amritsar
*Arunachal Pradesh	
Assam	Bongaigaon, Tinsukia, Dibrugarh, Silchar
Bihar	Jehanabad, Jamalpur, Muzaffarpur, Aurangabad, Arrah
Chhattisgarh	Ambikapur, Durg, Bhilai, Raipur, Korba
Goa	Mapusa, Margao, Mormugao, Vagator
Gujarat	Jetalpur, Deesa, Vadodara, Surat, Ahmedabad
Haryana	Faridabad, Gurgaon, Panipat, Narnaul
Himachal Pradesh	Dharampur, Solan, Mandi, Manali
Jharkhand	Medininagar, Ramgarh Cantonment, Ranchi, Dhanbad
Karnataka	Ranebennur, Bagalkote, Gangavathi, Mysuru
Kerala	Pala, Aluva, Kochi, Kozhikode
Madhya Pradesh	Nagda, Datia, Jabalpur, Gwalior, Indore
Maharashtra	Pune, Nagpur, Nashik, Kolhapur, Aurangabad
**Manipur	Imphal, Laipham Siphai
**Meghalaya	Shillong
**Mizoram	Aizawl
**Nagaland	Dimapur, Kohima
Odisha	Baripada, Bhadrak, Balasore, Rourkela
Rajasthan	Hanumangarh, Beawar, Tonk, Kota
**Sikkim	Gangtok
Tamil Nadu	Nagapattinam, Ambur, Vaniyambadi, Madurai
Telangana	Siddipet, Warangal, Khammam, Jagtial
**Tripura	Agartala
Uttar Pradesh	Shikohabad, Ghaziabad, Kanpur, Agra
Uttarakhand	Rudrapur,Haridwar, Roorkee, Rishikesh
West Bengal	Kharagpur, Durgapur, Bangaon, Asansol

*States did not have enough geolocations hence had to be discounted for the purpose of this study

**States were treated as a collective entity due to limited no. of geolocations within these states

To ensure that there are no residual cookies or any persistent data stored by Google after every query which could interfere with our result, we have used Google Chrome in incognito mode and added relevant code to delete cookies after every search. Other implementation details can be found in our GitHub Repository [3]. The links are stored in the Postgres database as an array of links, preserving the order in which they appear on the web page as the relevance of the order would become apparent as explained in the

Table 2: Few examples of search terms and different tags associated with them

Search Term	Tags
Covid Conspiracy	#controversial
Indira Gandhi death cause	#controversial, #notable_entities
China Covid Origin	#controversial
Best Hospitals	#medical
Aftab Poonawala	#notable_entities
Justice Loya case conspiracy	#controversial
Covid Origins	#controversial
Easy Jobs	#employment_economy
Best Jobs	#employment_economy
Best Hospital in India	#medical
Petrol Price	#employment_economy
Best Share to Buy	#employment_economy
LGBTQ Rights	#controversial
Unemployment Allowance	#employment_economy
Covid Deaths	#medical

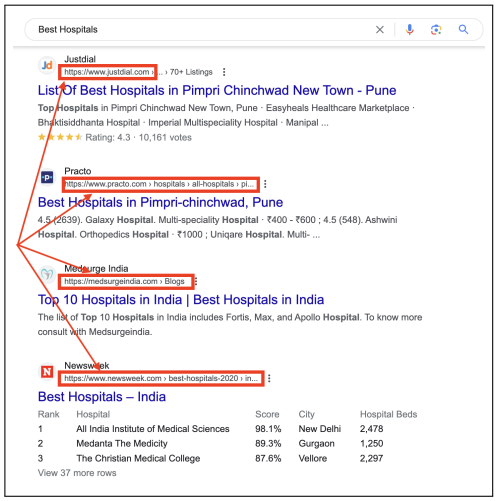


Figure 1: Exact form of data scraped

subsequent section. The output of this phase is a set of links for every search term for every geolocation.

3.3.2 ***Processing of the collated data using the RBO metric.*** As input to this phase, we have ordered sets of links for every search term across every geolocation. Therefore for every search term, the set of links for different geolocations needs to be compared to quantify and analyze the personalization observed across the different geolocations. So the problem reduces to comparing two sets. Now, instead of comparing every pair of links for each location individually, we opted for a more efficient approach. We chose a reference city against which we will compare all other cities.

For the first set of locations, i.e ones used for National-level study, we have chosen the capital of India, New Delhi, as a base city, and

Table 3: Examples of base cities for some states

State	Base City	Other cities
Uttar Pradesh	Lucknow	Shikohabad, Ghaziabad, Kanpur, Agra
Odisha	Bhubaneswar	Baripada, Bhadrak, Balasore, Rourkela

sets of all other cities (i.e. capitals of all the states of India) is compared to the set of links obtained for New Delhi. For the second set of locations, i.e ones used for State-level study, the state capital is chosen as the base city and the set of links for all other geolocations of that particular state are compared to the set of links scraped for the state capital (Refer Table 3).

In this research, our primary objective centered on quantitatively analyzing two lists of web links, with the possibility of these sets containing differing or reorganized web links. Consequently, we reviewed several similar studies [5, 9, 13] and discovered that they employed a range of metrics, including the Jaccard similarity coefficient, the minimum edit distance, the Kendall Tau rank correlation coefficient, and the Rank-Biased Overlap (RBO) measure.

After careful consideration, we determined that the Rank-Biased Overlap (RBO) [29] metric was the most suitable for our research objectives. This decision stemmed from the fact that RBO does not require the lists to be conjoint (same element in both lists) and incorporates weighting, assigning varying importance to different positions. This weighting feature is crucial for our study because the arrangement of web links on a page plays a pivotal role. It is observed that individuals are more inclined to click on links located at the top of the web page rather than those at the bottom. RBO takes values in the range [0, 1] where 0 means disjoint and 1 means identical ranked lists. In conclusion, RBO emerges as a more robust metric for assessing the similarity of non-conjoint ranking lists and offers the flexibility to be adjusted to prioritize the top positions.

We have used the RBO module [4] available through python package installer (pip). After the RBO values are calculated for every geolocation with the base city, they are stored in a Postgres database. This marks the end of this phase.

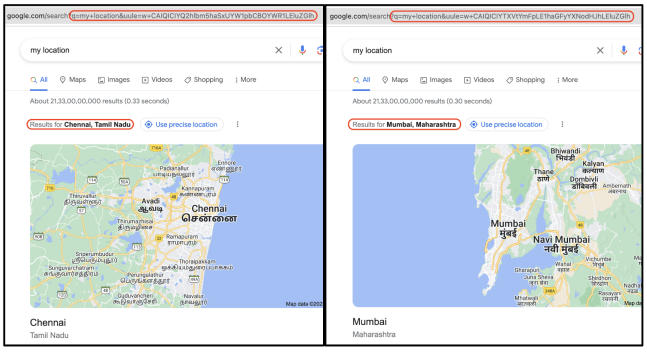


Figure 2: Modifying 'uule' parameter to change geolocation

3.3.3 ***Visualising the raw numeric data for analytical ease.***

In the final phase, the collected data is presented using various methods. Out of the wide variety of data visualization methods, we have chosen two standard plots and one non-standard which is unique to our study but provides the best visual synopsis of our results in comparison to all other visualization techniques available.

The standard ones include a boxplot and bar graphs. The boxplot gives information about the spread of the entire data, as well as the maximum and minimum RBO values obtained during our analysis. Bar graphs are a simple way to visualize the RBO values in a sorted manner allowing for quick insight into our observations. The box plot is included in this paper however due to space constraints, the bar graphs can be found in the 'complete data' folder of our GitHub repository [3].

The non-standard approach is visualization through map representation. This proves to be a novel approach to visualizing disparity in search results between the various states of India and enables us to infer meaningful conclusions with minimal effort. It includes aggregation and subsequent correlation of the RBO values obtained for each state to a particular shade of the color red. This color then is plotted on the map of India. The resulting heatmap gives us a unique way of visualizing the differences in search results by region. India polygon shapefiles were obtained from [20].

Therefore, for the first set of locations, we have calculated the mean deviation of the state capitals from the country's capital over many search queries and have plotted a heatmap over the map of India.

For the second set of locations, we have created another interpretation of this map where each state is treated as its own entity. To recall, we had used the state's capital as the base city and had picked 3-5 other cities. Therefore, we calculate the mean deviation of the other cities of the state from the state's capital and plot the results on a heatmap over the map of India.

4 OBSERVATIONS AND DATA ANALYSIS: NATION-WIDE STUDY

This section presents the data obtained from the first set of locations, i.e., the variation of each state with respect to a single reference city, New Delhi, and subsequently discusses the inferences that can be drawn from the results.

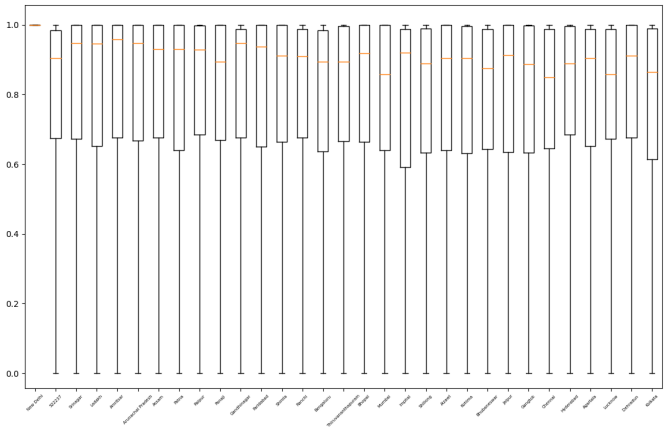
The first step involved the creation of a box plot illustrating the aggregated RBO value for multiple queries, organized by state, as depicted in Figure 3. While the length of the boxes consistently spans from 1 to 0.6 for all the states, the substantial length of each box indicates the presence of bias in each state.

So, it is worth checking individual search terms for each state as some interesting observations can be made since the box plot indicates bias in one or more search terms for all the states. Now, let us examine some individual search terms that had interesting results.

4.1 Search Term: Marijuana Legal

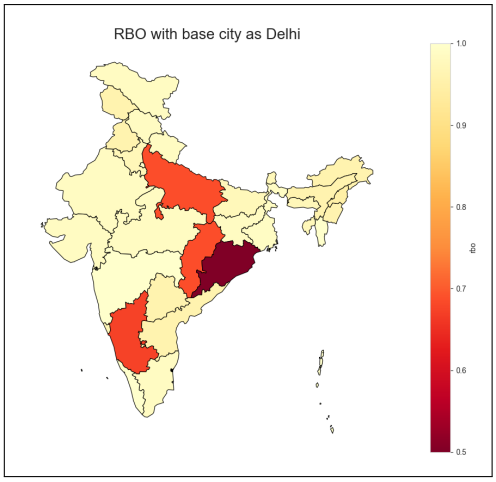
This search term shows a fair amount of variance in selected cities, as observed from the heatmap (Figure 4).

Figure 3: State-wise Aggregation of RBO Values



Note X-axis represents the states, and the Y-axis indicates a box plot for each state, which aggregates all the RBO values measured for every search term.

Figure 4: Heatmap for search term: Marijuana Legal



We can see that although several cities have results similar to New Delhi, few locations are very different. These include Karnataka, Odisha, Chhattisgarh, and Uttar Pradesh.

India is a country where marijuana is banned in every state. Thus, the presence of bias in certain states is unexpected and might represent varied opinions among the people, as its legalization is a controversial subject.

4.2 Contrasting search terms: 'Buy Samosa' and 'Buy Solar Panel'

Another interesting observation is that the more popular and local a product is, the more localized the results are. For example, we ran two queries 'Buy Samosa' and 'Buy Solar Panel'. The heatmap for both is shown below in Figure 5. We know that samosa is more popular and a local purchase item than a solar panel. So based on our hypothesis, we should get more localized results for the

samosa search term. It can be seen that the colour of the heatmap

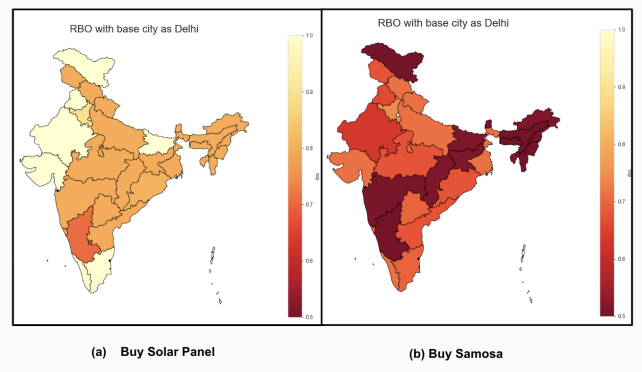


Figure 5: Contrasting 'Buy Samosa' and 'Buy Solar Panel'

for the query 'Buy Samosas' is red and more varied compared to the heatmap for the query 'Buy Solar Panels'. Therefore, the results are more varied, and our hypothesis is confirmed. This can be further confirmed using other valid pairs like 'Restaurants near me' and 'Tractor near me'. In this case, it is expected for the more popular search term to have more localization though both the search terms need localization.

4.3 Search Term: Agriculture

Another search term where variance was observed over most parts of India is "Agriculture". Observe Figure 6.

As observed, there is a small deviation in almost all the states of India. The highest deviation is observed in Kerala, possibly due to the big difference in the kind of agriculture in these states (kind of crops grown, government schemes, and type of soil). Same can be inferred for all such states which have visible anomalies. Furthermore, it is important to highlight that the website *https://dbtagriculture.bihar.gov.in/* appears for multiple

Figure 6: Heatmap for the search term 'Agriculture'

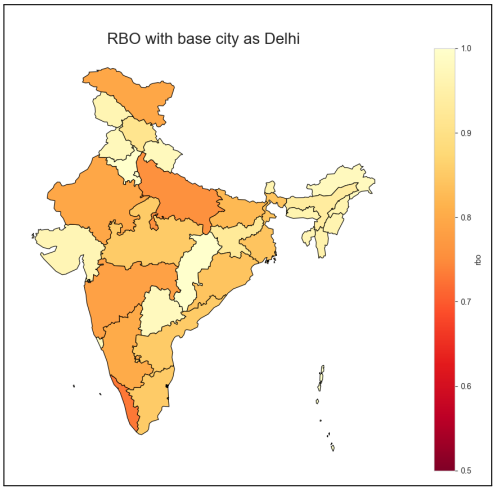


Table 4: Tags and complete set of queries associated

Tag	Queries
Medical	Best Hospitals, Best Hospital in India, Covid Deaths, Covid Rules
Employment_Economy	Easy Jobs, Best Jobs, Petrol Price, Best Share to Buy, Unemployment Allowance, Tech Jobs, Police, Jobs That Pay the Most, Domestic Help, Study Abroad, Engineering College, Inflation, Agriculture
Expected_Local	Engineering College, Buy Solar Panels, Domestic Help, Buy Samosas, Legal Drinking Age

geolocations, as illustrated in Figure 7. This URL corresponds to the Agricultural Department of the Bihar State Government, suggesting a potential bias towards agriculture in the state of Bihar.

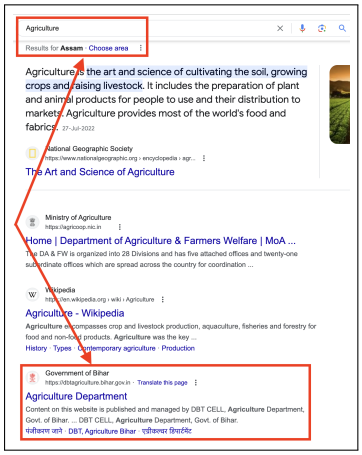


Figure 7: Anomalous result for query 'Agriculture' in Assam

4.4 Aggregation of Search Terms on the basis of associated tags

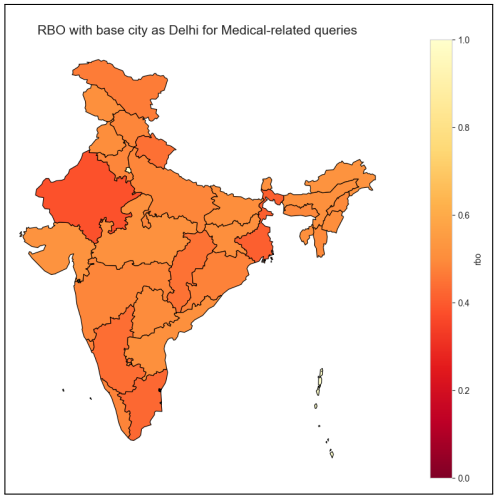
During the process of formulating the search terms, we categorized them by assigning tags, as outlined in the methodology section. In this subsection, we aggregate the RBO values for the popular search terms based on their respective tags and represent the data in the form of a heatmap. The table containing the list of queries associated with each tag is presented in Table 4.

4.4.1 Tag: Medical.

First, we have the search terms tagged as "Medical". The map created after compiling RBO values of medical-related search terms is given below (Figure 8).

In general, it is observed that while there is sufficient variance, as expected, in such search terms, results related to Mayo Clinic's top 10 hospitals in the USA are shown very frequently, and thus the search terms are never completely localized.

Figure 8: Heatmap for the tag '#medical'

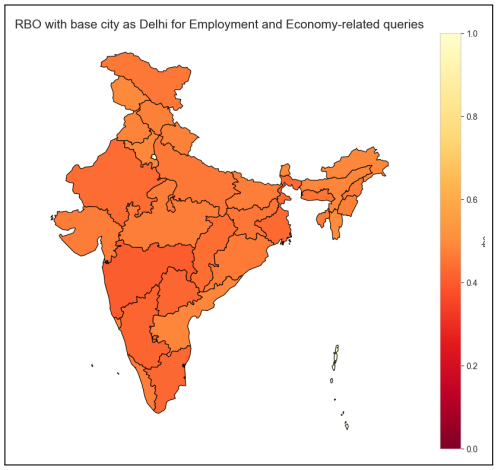


4.4.2 Tag: Employment_Economy.

Second, we have the search terms tagged as "Employment_Economy." The map created after compiling RBO values of Employment and Economy related queries is shown in Figure 9.

It can be seen that the variance in RBO is largely adequate and relatively uniform, indicating a mixture of localized and some global results that are shared with the reference city, New Delhi.

Figure 9: Heatmap for the tag '#employment_economy'

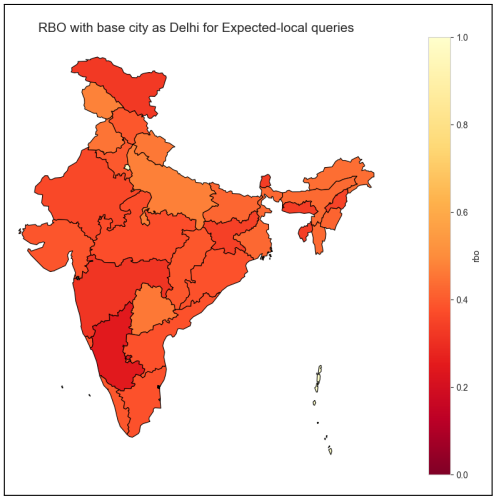


4.4.3 Tag: Expected_Local.

Finally, we have search terms tagged as "Expected_Local". This shows a better picture of the degree to which a state shows differences in Google search results as it shows those queries which were expected to have local results but ended up having global results. The map created after compiling RBO values of Expected-Local related queries is shown in Figure 10.

We can observe that states like Jammu and Kashmir, Karnataka, and a few others have a remarkably darker shade of red. This indicates the presence of considerably different local entities for these states. Perhaps the lifestyles of individuals in these states are notably different when compared to the country's capital New Delhi.

Figure 10: Heatmap for the tag '#expected_local'



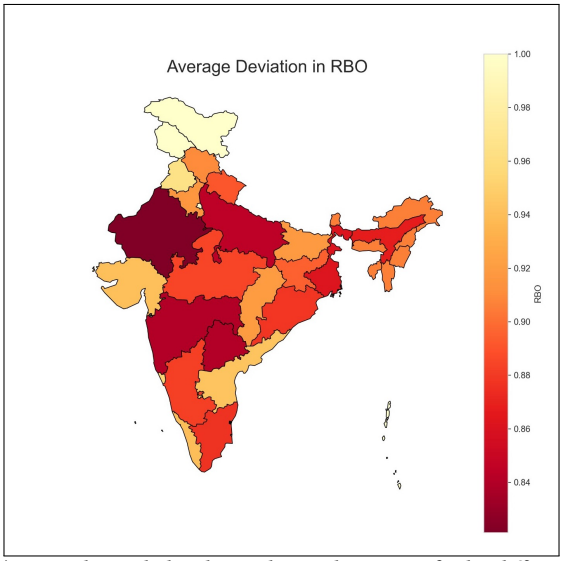
5 OBSERVATIONS AND DATA ANALYSIS: STATE-WIDE STUDY

This section presents the data obtained from the second set of locations, i.e. the variation of each chosen city of the state with respect to the state's capital, and subsequently discusses the inferences that can be drawn from a few of the interesting results. After running the appropriate queries for various cities in each state, we plotted a map where the gradient of each state shows how much within-state variation has been measured (Refer Figure 11).

India is highly diverse and has various cultures. Several cities in India have populations higher than countries. Due to this, it is expected that there will be significant personalization in Google search results, especially in metropolitan cities where people, in general, are cosmopolitan. Some interesting observations are:

- It is observed that Rajasthan seems to have the largest within-state variation. This can be attributed to it being the largest state in India. In general, it can be observed that the bigger the state, the higher the within-state variation.
- Furthermore, Maharashtra also has a high within-state variation. This observation is reinforced by the search results [12] (for the query 'Legal Driving Age') and [27] (for the query 'National Farmer's day') appearing only in Mumbai, while the other cities of Maharashtra do not show these particular links. This might be due to the fact that Mumbai's cosmopolitan population & diverse cultures are distinct from the rest of Maharashtra.
- To specify yet another supposed anomaly, for the search term 'police' for the geolocation 'Shikohabad, Uttar Pradesh', results related to the police department of Rajasthan (*https://www.police.rajasthan.gov.in/*) are observed. This may be attributed to the

Figure 11: Heatmap for State-wise Study



*Note: The scale has been changed to magnify the differences observed

proximity of Shikohabad to Rajasthan and the limited prominence of Shikohabad as a city.

This concludes our observations and findings for a state-wide study for quantifying personalization induced by Google's search engine.

6 FUTURE RESEARCH DIRECTION AND OPEN CHALLENGES

Our research provides a clear starting point for studies regarding the impact of search engine results on the users' behavior. The methodology paves way for conducting extensive research in the realm of search engine personalization. Right from web link scraping, the usage of RBO for comparative analysis, and visualizing obtained results as a heat map of India, all make the methodology a solid foundation for future research in this field. Some extensions to the study could include analysing topics of national interest.

For instance, a news piece of a national scale should not differ with changing locations. It may indicate some form of media control and censorship. A differing RBO can be further investigated to verify if there is a difference in facts. Another possible scope for this data can be for product-based companies to alter their offerings and cater to the different biases that people in various locations may have.

This study can also be conducted for other notable search engines like Bing and DuckDuckGo. In fact, the latter would be an intriguing candidate as it claims to be free of all cookies and should show considerably less variation.

We have conducted this study taking a single city as the base city. Pair-wise research for each state and city can offer invaluable insights into Google's search results at the city and state levels.

Our method of spoofing location can result in high variation when replicating results. A solution can be volunteering people from different places to provide the actual results from their locations for more accurate results.

Although RBO is much better than previous methods of comparing two lists, it lacks precision in highlighting the specific differences between the two lists.

Our list of queries is extensive but not exhaustive. Different search terms and alternative search queries might yield distinct outcomes. It can be challenging to come up with a list of queries that is large enough to cover most of the searches done by people and also be diverse enough to cover all the different locations in India.

7 CONCLUSION

Google is the most used search engine in India. It is the go-to platform for information when one wants to study across a multitude of domains. In our research, we have aimed to analyze Google search results, specifically the personalization in Google search results due to geolocation. We have carefully curated a list of queries to facilitate further analysis and draw explanations to justify the presence of personalization.

Our research is the first to isolate location as the sole factor while the existing literature addresses personalization using several other factors like age, gender, location and so on. Moreover, India is a diverse country with varying cultures across various regions. This remarkable diversity is not only evident across different states, but also across cities within the same state. Location as a factor hence provides a great representation to capture how these changing demographics affect search engine personalization. Hence we have collected the data for each of the Indian states and analyzed the results for every pair. Furthermore, we have used this data to compare various cities within each state. Thus we have considered both scenarios and divided our data collection and analysis into two phases to address the same. After collecting the relevant data, we analyzed and drew various conclusions, some of which are the following:

- Within-state personalization is high in larger states or cities with cosmopolitan populations. This is evident as seen in the search results obtained for states like Rajasthan and Maharashtra with their large size and diverse population.
- Items of personal use, readily available within a local context, such as the popular food item, Samosa, tend to yield search results that are predominantly localized in nature. Conversely, niche products like Solar Panels often produce search outcomes that exhibit a more globalized scope. This suggests that queries involving locally prevalent items demonstrate a heightened degree of personalization in search results.
- It was observed that certain queries show a global result where localization is expected as seen in Medical-related search terms. Moreover, presence of Bihar Agricultural Department in search results of other states was unexpected.

After analyzing the collected data and drawing appropriate conclusions, we can attest that personalization is pervasive across Google. While the extent of personalization varies, it is consistently present. Personalization can often be beneficial. However, it can also lead to the Filter Bubble Effect, which makes understanding personalization all the more significant.

REFERENCES

[1] 2009. *WWW '09: Proceedings of the 18th International Conference on World Wide Web* (Madrid, Spain). Association for Computing Machinery, New York, NY, USA.

[2] Leonardo Andrade and Mário Silva. 2006. Relevance Ranking for Geographic IR.

[3] Pranav Chatur. 2022. GitHub - DarkMenacer/Google-Geolocation-Bias: Tool to scrape web links for different queries from different cities across the world, all from your home — github.com. https://github.com/DarkMenacer/Google-Geolocation-Bias. [Accessed 13-Jul-2023].

[4] Changyao Chen. 2021. *rbo 0.1.3.* Retrieved Jan 31, 2023 from https://pypi.org/project/rbo

[5] Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. 2007. A large-scale evaluation and analysis of personalized search strategies. *16th International World Wide Web Conference, WWW2007*, 581–590. https://doi.org/10.1145/1242572.1242651

[6] Sharad Goel, Jake Hofman, and M. Sirer. 2021. Who Does What on the Web: A Large-Scale Study of Browsing Behavior. *Proceedings of the International AAAI Conference on Web and Social Media* 6, 1 (Aug. 2021), 130–137. https://doi.org/10.1609/icwsm.v6i1.14266

[7] Google Ads API 2023. *Geo targets.* Retrieved Nov 27, 2023 from https://developers.google.com/google-ads/api/data/geotargets

[8] Government of India, Ministry of Finance, Department of Expenditure 2015. *Re-classification/Upgradation of Cities/Towns on the basis of Census-2011 for the purpose of grant of House Rent. Allowance(HRA) to Central Government employees.* Retrieved May 27, 2017 from https://doe.gov.in/sites/default/files/21-07-2015.pdf

[9] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring Personalization of Web Search. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13)*. Association for Computing Machinery, New York, NY, USA, 527–538. https://doi.org/10.1145/2488388.2488435

[10] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. 2007. Demographic prediction based on user's browsing behavior. *16th International World Wide Web Conference, WWW2007*, 151–160. https://doi.org/10.1145/1242572.1242594

[11] IAMAI. 2022. *KANTAR, Internet in India 2022.* Retrieved Sept 2, 2022 from https://www.iamai.in/sites/default/files/research/Internet%20in%20India%202022_Print%20version.pdf

[12] indiacode. 2022. *Legal Driving Age.* Retrieved Nov 2, 2022 from https://www.indiacode.nic.in/show-data?actid=AC_CEN_30_42_00009_198859_1517807326286&orderno=6#:~:text=(1)%20No%20person%20under%20the,the%20age%20of%20sixteen%20years

[13] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. 2015. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Proceedings of the 2015 Internet Measurement Conference* (Tokyo, Japan) *(IMC '15)*. Association for Computing Machinery, New York, NY, USA, 121–127. https://doi.org/10.1145/2815675.2815714

[14] Fang Liu, Clement Yu, and Weiyi Meng. 2002. Personalized Web search by mapping user queries to categories. *International Conference on Information and Knowledge Management, Proceedings*, 558–565. https://doi.org/10.1145/584792.584884

[15] Maryam Mohsin. 2022. *GOOGLE SEARCH STATISTICS YOU NEED TO KNOW IN 2023.* Retrieved Dec 31, 2022 from https://www.oberlo.com/blog/google-search-statistics

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

[17] James Pitkow, Hinrich Schütze, Todd Cass, Rob Cooley, Don Turnbull, Andy Edmonds, Eytan Adar, and Thomas Breuel. 2002. Personalized Search. *Commun. ACM* 45, 9 (sep 2002), 50–55. https://doi.org/10.1145/567498.567526

[18] Population density 2022. *Population Density.* Retrieved Sept 27, 2022 from https://www.indiacensus.net/density.php

[19] Alexander Pretschner and Susan Gauch. 1999. Ontology based personalized search. 391 – 398. https://doi.org/10.1109/TAI.1999.809829

[20] Princenihith. 2020. Maps with Python. https://github.com/Princenihith/Maps_with_python. Accessed on 28 May 2023.

[21] Feng Qiu and Junghoo Cho. 2006. Automatic identification of user interest for personalized search. *Proceedings of the 15th international conference on World Wide Web*, 727–736. https://doi.org/10.1145/1135777.1135883

[22] Kevin Richard. 2020. *Geolocation: The Ultimate Tip to Emulate Local Search.* Retrieved Nov 16, 2022 from https://moz.com/blog/geolocation-the-ultimate-tip-to-emulate-local-search

[23] Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Implicit User Modeling for Personalized Search. *International Conference on Information and Knowledge Management, Proceedings* (10 2005). https://doi.org/10.1145/1099554.1099747

[24] Jian-Tao Sun, Hua-Jun Zeng, Huan Liu, Yuchang Lu, and Zheng Chen. 2005. CubeSVD: A novel approach to personalized Web search. *Proceedings of the 14th International Conference on World Wide Web*, 382–390. https://doi.org/10.1145/1060745.1060803

[25] Bin Tan, Xuehua Shen, and ChengXiang Zhai. 2006. Mining long-term search history to improve search accuracy. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2006, 718–723. https://doi.org/10.1145/1150402.1150493

[26] Jaime Teevan, Susan Dumais, and Eric Horvitz. 2005. Personalizing search via automated analysis of interests and activities. 449–456. https://doi.org/10.1145/1076034.1076111

[27] vikaspedia. 2022. *National Famers day.* Retrieved Nov 2, 2022 from https://vikaspedia.in/agriculture/agri-directory/important-days/national-farmers-day#:~:text=National%20Farmers'%20Day%2C%20also%20known,in%20India%20on%2023%20December.&text=Your%20browser%20can't%20play%20this%20video

[28] Dr.N.Srinivasan V.Raju. 2021. An Efficient Analysis of Web Search Personalization Using Fuzzy Based Approach. *Turkish Journal of Computer and Mathematics Education* 12, 12, Article 5 (March 2021), 3753-3759 pages. https://www.turcomat.org/index.php/turkbilmat/article/view/8153/6365

[29] William Webber, Alistair Moffat, and Justin Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.* 28, 4, Article 20 (nov 2010), 38 pages. https://doi.org/10.1145/1852102.1852106

[30] Bo Yu and Guoray Cai. 2007. A query-aware document ranking method for geographic information retrieval. 49–54. https://doi.org/10.1145/1316948.1316962

Response to Reviewer Comments

48: Google Search in India: Unveiling the Geo-Personalized Web

—----------

We are immensely grateful to the esteemed reviewers for their invaluable and constructive feedback. In this attached document, we have carefully analyzed and addressed the concerns raised by each of the reviewers. The insightful suggestions provided have profoundly influenced us to refine our work's comprehensibility.
Here are our responses to the revision requests put forth by each of the reviewers.

—----------

Reviewer 1

—----------

Overall Concern: The paper is an empirical study, but doesn't rationalize the choices and how they contribute to their conclusion. There seems to be some significant cherry-picking.

Response:
We acknowledge that the presentation of our findings may have conveyed an impression of data selectivity. The queries chosen for our study were deliberately diversified, to include a wide spectrum of subject matters, including healthcare, education, employment, politics, and climate. (list of queries). We have chosen these queries because they pertain to a substantial demographic, without any particular bias or motive. We carefully examined all these queries and created combined maps that considered every query.
We have visualized the results for the complete data through box plots, and our analysis has revealed a discernible degree of personalization influenced by variations in geolocation.
Hence, we have conducted a separate discussion for tags whose associated queries exhibited the most prominent influence on the aforementioned personalization.
Alongside this, we made an effort to explain the possible reasons for these results. We acknowledge that this might have seemed like we were being selective, which was not our intention. We've made adjustments to clarify how we selected the queries for emphasis and modified the conclusion accordingly **(Page 8, Section 7)**.

—----------

—----------

Revision Request 1: Explain relevance of RBO, and some details regarding it too. Also the same thing about scraping method

Response:
We acknowledge that there have been some slight gaps in our explanation regarding the relevance of RBO in our study along with the scraping method. Keeping this in mind, the paper has been revised (**Page 3 Section 3.3.1, Page 4 Section 3.3.2**) to provide a clearer understanding of the same.

—----------

Revision Request 2: While the paper presents interesting findings, the analysis could be further elaborated to provide a deeper understanding of the reasons behind specific observations and anomalies.

Response:
Evaluating the reason behind the specific observations would require us to provide our subjective opinion on the internal workings of Google's search engine algorithm. We felt that this was not the purpose of this study and wanted it to remain as objective as possible by only studying the similarity (or dissimilarity) of these search results.
Any individual or entity seeking to conduct an analysis to explore the underlying factors contributing to these observations will discover our research to be a valuable launching point. Our code is entirely open-sourced under MIT License, and we are readily available to offer assistance and support in any capacity that may be required.

—----------

Revision Request 3: The paper could benefit from a broader discussion of how its findings relate to the existing body of research on search engine personalization, both globally and within India.

Response:
The study, in the process of quantifying the variations and presenting them in a way that feels most intuitive to the reader, contributes to the existing literature in the following ways.

The existing literature primarily addresses personalization using several factors based on the user's profile like age, gender, and their location to mention a few. Our finding is the first to isolate location as the sole factor for personalization. Moreover, India is a diverse country with varying cultures across various regions. Location, as a factor, provides a great representation to capture how these changing demographics affect search engine personalization.

Further, the methodology itself paves the way for conducting extensive research in the realm of search engine personalization. Right from web link scraping, the usage of Rank-Biased Overlap (RBO) for comparative analysis, and visualizing obtained results as a heat map of India, all make

the methodology a solid foundation for future research in this field. Our code is open-source and freely available on GitHub under MIT License.

—----------

Reviewer 2

—----------

Revision Request 1: While the study's methodology and execution are commendable, some aspects of the findings were somewhat anticipated. Given the growing awareness of the impact of geolocation on personalized search, it's not entirely surprising that the study's results aligned with certain expectations.

Response:
Our study primarily deals with quantifying the variations observed when particular queries are searched across India (as compared to the country's capital, New Delhi). The inferences derived from the observations do not aim to propose radical explanations for the supposed variations. Therefore, throughout our study we have been conscious enough to put forward the observations in the most faithful manner.

Any anomaly observed can further be explored by analyzing the specifics related to the search term and location explicitly but remains an extension of the paper rather than being the premise of it.

—----------

Revision Request 2: While the decision of using RBO is justified, it lacks precision in highlighting specific differences between the two lists. Thus RBO is not sufficient to understand how exactly personalized page ranks affect user behavior.

Response:
Accurately capturing how users are affected by personalized page ranks would require a few additional steps to the methodology proposed in the paper. Thus we believe that our research paves the way for conducting a broader study, in which aspects related to a specific individual can be tinkered with while keeping the basis of comparison to be a standard impersonalized search.

—----------

Revision Request 3: The search queries could have also included cases where the search results should have been the same across geo-location, for example, topics of national importance.

Response:

The search terms presented in the paper are just a subset of the actual list of search terms used for the study. In lieu of space, we chose to mention only a few of them in the paper but we have stored all of them in our database (the files which are provided along with our code).
With that said, we do believe that we could have included more national-level topics, but we instead chose to include search terms of a variety of domains to try and capture how personalization varies across other domains (as evident in the 3 maps presented in **section 4.4(Page 6)** of the paper).

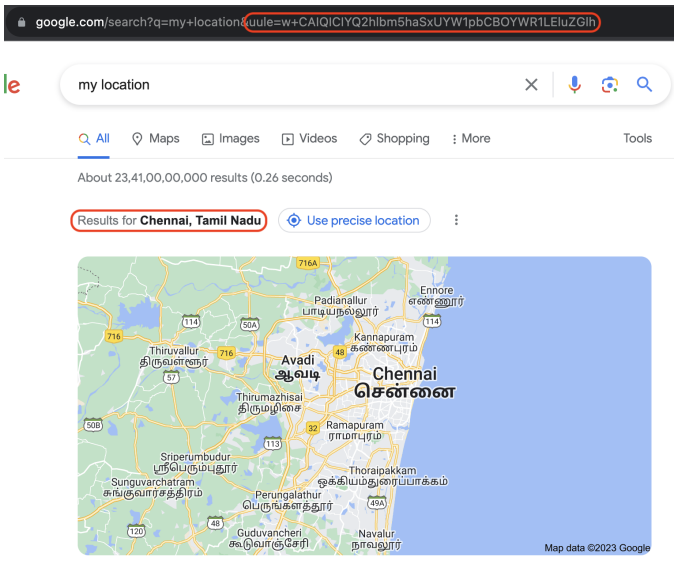
—----------

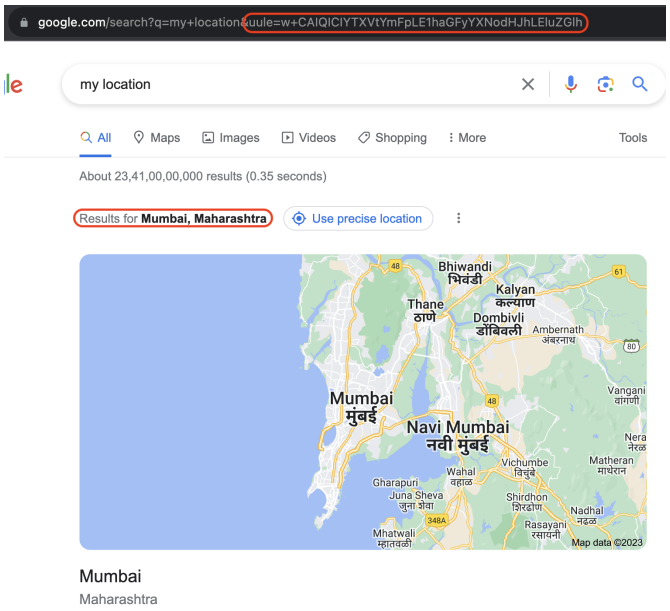
Reviewer 3

—----------

Revision Request 1: The authors assume that setting just one parameter in the URL makes the search localized. More concrete evidence should be shown to support this. Without this, the finding could be considered inconclusive.

Response:
Obtaining localized results was of utmost importance in our study. To do so, we use the URL parameter ‘uule’ (query parameter used to encode a place or an exact location).





To strengthen the methodology in contention (filtering results based on geolocation using 'uule,') we would like to state the following sources which also mention that such a strategy is employed widely to filter results based on geolocation:

1. https://moz.com/blog/geolocation-the-ultimate-tip-to-emulate-local-search
2. https://blog.linkody.com/seo-local/uule-2
3. https://valentin.app/uule.html

In the revised document, we have resolved this matter in two ways

First, we have added an image (**Figure 2 on Page 4**), similar to the one above, contrasting the different values of uule parameter for two different locations, namely Mumbai and Chennai. This image serves the purpose of satisfying readers about the change of location while displaying the results.

Second, we have not only added reference to the above links but also provided clearer explanations to show how the URL parameters are being tampered with in **section 3.3.1 (Page 3)** (Scraping the data from Google Webpages) of the paper to further substantiate our stance to obtain desired results.

—----------

Revision Request 2: The authors did not pursue to identify what factors caused the bias in the results.

Response:
Our goal has been to stay as objective in our study as possible. With the data we have gathered, we could only speculate the reasons for the anomalous results. We have provided some explanations (the states of Rajasthan and Maharashtra showing higher within-state variation due to their size and cosmopolitan population respectively) but held off from giving a strong opinion about the workings of the Google search engine.

—----------

Revision Request 3: Now, since all the queries have been issued from the same machine and from the same location, some bias may inherently creep in.

Response:
While it is true that these queries were issued from the same geographic location i.e. Jaipur, the task of data collection was distributed across several machines. No single machine has issued all the queries used in the study.
All efforts were made to ensure that inherent bias was kept to a minimum through the use of undetected-chrome-driver, incognito mode, as well as deleting cookies after each search. However, we concede that there might be some unknown variables affecting our results which would be ubiquitous in all previously conducted studies on this matter.

—----------