Detecting Geographical Bias in Search Engine Queries

Raja Iqbal
Department of Statistics,
Stanford University, Stanford, CA
riqbal@stanford.edu

Abstract

We propose an approach to detect bias in web search engine queries towards a particular geographical region. The term bias here refers to the fact that some search engine queries will be more likely to occur in certain geographical location than other queries. We have used a likelihood ratio (similar to KL divergence) based technique to obtain the bias of a given query toward a particular geographical region. The results show interesting socio-economic and political patterns across different states of the United States.

Keywords: Unsupervised Learning, KL Divergence, Bayes Rule, Query Classification.

1. Introduction

Knowing if a query is more likely to occur in certain geographical region can be extremely useful for various reasons. The queries coming from a geographical region can give marketers an insight into the preferences of the population of the region. The information can be used to market new products in these regions. Additionally this information could be used as one of the features during behavioral targeting for search ads [3].

In this study, we propose a method to understand if queries issued by search engine users have an inherent bias due to their geographical location. Our premise is that the web queries are representative of the behavior of population residing in a given geographical region. We currently limit our analysis to the United States of America. We estimate the bias of certain queries towards a particular geographical region using a very simple KL Divergence like criteria. For instance, what queries are likely to occur in the state of WA compared the rest of the United States.

2. Detection of Geographical Bias Using Likelihood Ratio

2.1. Data Collection

We used search impression and click-through data from a commercial search engine for the month of April to conduct this study. The log dataset records all users' search and click behavior. Queries to the web vertical were chosen from users in the US market with language preference set to English. Any search records suspected of being bots were filtered out from the dataset before conducting the experiment. Records are marked as bot based on various criteria like query entropy, number of queries per session, pagination click frequency, ad click frequency, time interval between successive queries etc. Explaining what constitutes a bot is beyond the scope of this work. Interested readers are encouraged to read a good survey paper [2] on web bot detection to understand how bot traffic is handled by commercial search engines.

The total number of users in the dataset was close to 10 million with a total of more than 1.4 billion queries of which 264 million queries were distinct queries. A typical query-click sequence in the logs is shown in Figure 1 below.

User ID	Time Stamp	Event Type	Event Value
User 1	20071205110843	QUERY	KDD
User 1	20071205110844	CLICK	www.kdd.org
User 1	20071205110845	CLICK	www.kdd2008.com

Figure 1: Visualization of a few rows of search logs

2.2. Query Preprocessing

For each query in the dataset, twenty five common stop words were removed. The plural forms of data and capitalization were also removed. For instance, a query *'Buying cars in the Seattle area'* would be normalized to *'buying car seattle area'*.

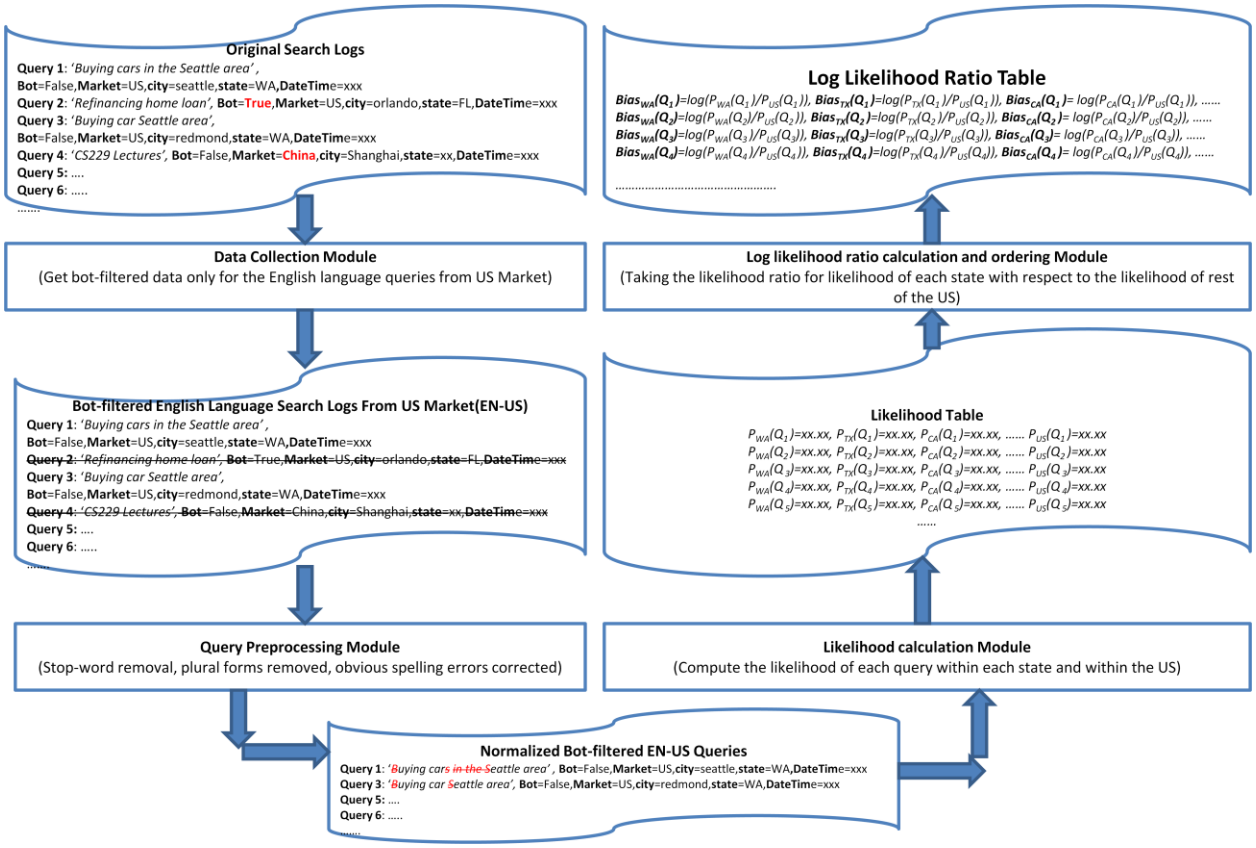

Figure 2: Overall flow of the learning process. Note how the records marked as bot and not belonging to the US market are filtered. Also note the normalization of the queries.

2.3. Computing the Query Likelihoods

After performing the bot-filtering, market detection and normalization on the input queries, we computed the likelihood each query. To compute the likelihood of the queries we further filtered the normalized queries and removed the queries whose frequency of occurrence is less than certain threshold. It is very common for search engine logs to contain noise like punctuations, repeated characters and other garbage characters in the query logs. Thresholding based on the frequency of occurrence ensures that any noise does not skew the query distribution.

For each query Q_i the likelihoods of the query in the state of WA and within the US is computed by

$$P(Q_i|state = WA) = \frac{frequency\ of\ occurence\ of\ Q_i\ in\ WA}{tnumber\ of\ queries\ in\ state\ of\ WA}$$

Similarly, the likelihood $P(Q_i)$ is computed by

$$P(Q_i) = \frac{frequency\ of\ occurence\ of\ Q_i\ in\ United\ Stated}{number\ of\ queries\ in\ United\ Stated}$$

For the same query Q_i, the bias toward the state of Washington is computed as

$$Bias(Q_i|state = WA) = \log\frac{P(Q_i|state = WA)}{P(Q_i)}$$

This bias will be high when a query is more likely to occur in WA compared to rest of the US. Similarly the number will be small if the query is less likely to occur in WA than US. Once we have computed the likelihood ratios for all the queries within a state, the queries are sorted by likelihood ratio. The queries with the highest value of the log likelihood ratio are considered to be the ones that are highly likely to occur within that state. Likewise queries with the lowest values of the likelihood ratio are the ones that are least likely to occur in the given state. We computed the bias of queries for ten US states as shown in Figure 3.

A natural question that arises is why is just the frequency of occurrence of respective queries not sufficient and why do we take the ratio of the likelihoods in the state and the rest of the United States as an indicator of bias. We address this question by observing that certain head queries are very popular all over the US, irrespective of the geographical location of the origin of query. Queries like, 'youtube' and 'facebook' are always the queries among the highest frequency queries no matter what the geographical location is. Considering only the frequency of occurrence would put these queries among the top queries in each of the states and hence the queries most likely to occur in a given geographical location. For a query that is uniformly distributed among all the states of the US, the log likelihood ratio will be zero. For a query whose likelihood is higher in a given state than the rest of the US, the likelihood ratio will be a number greater than one and hence the log likelihood ratio will be a positive number. Similarly, for a query having a lower likelihood value in a given state than the rest of the US, the likelihood ratio will be a number less than 1 and hence the log likelihood ratio will be a negative number.

3. Results and Discussion

We processed one month of search logs from a major web search engine. The data was of the order of several hundred terabytes. Custom scripts were written to perform the query processing and calculation outlined earlier in this report. The scripts were run on a high performance computing cluster with several hundred machines.

Based on the first prototype that we implemented, we have observed interesting patterns in web search queries from the US market during the month of April 2009.

Some of the observations are discussed below

- *News Channel:* The news channel preferences observed in the logs is consistent with the expectations. For instance, Republican majority states clearly show a preference for Fox news as opposed to CNN which seems popular in Democrat majority states.

- *Social Networking Sites:* We also observed that population in a state has different preferences for social networking websites. For instance Michigan has facebook as a query that is strongly biased towards the state whereas myspace is biased away from MI.

- *Housing Market:* Zillow is one of the top queries in CA. This may be an indicator of something happening in the housing market in CA. Most likely people are taking advantage of the housing market and foreclosures in CA.

- *Travel Related Queries In Hawaii:* Hawaii seems to have an unusually high number of travel queries biased towards it which is expected.

- *Swine Flu:* While there are swine flu related queries all over the US, TX seems to have a higher likelihood of having this query. This is consistent with the observation that patient zero for swine flu was found in Mexico and the geographical proximity of TX and Mexico.

- *Politics:* Queries related to Gov. Sarah Palin dominate the top 20 queries in the state of Alaska. This can be explained by the fact the Sarah Palin is from Alaska and has been in news in recent past.

- *Lottery:* Queries related to lottery seem to be popular among many different states such as NY, CA and FL.

- *Unexplained Query Behavior:* We observed several queries that do not have a clear indication why the query is among the top queries. For instance, query '*mitchell's gourmet foods union*' is the query that is most likely to come from TX. After several web searches we could not find the correlation between this specific query and the state of TX. One possible explanation is that some search engine optimization company based in TX is working to boost the rank of some website by issuing the query again and again and our logic to detect bots failed to mark them as such.

Please refer to the demo attached with the submission email for more details on the most likely and least likely queries for each of the states.

4. Future Work

Based on the observations from the results we would like to pursue following items for improving the quality of this work

Removal of location names from queries during normalization: Currently there are several queries that have a location name co-occuring. "Craigslist <LocationName>" is one such examples. It would be interesting to see how the results change when we remove the location name and then obtain a distribution of queries. After removing the location term from the query,

"Craigslist Seattle" and "Craigslist SF Bay Area" would be mapped to the same normalized query "craigslist". This will ensure a more accurate comparison. Currently "Craigslist Seattle" and "Craigslist Boston" show at the top of the list of queries biased towards the state of WA and MA respectively. This may be due to the location component and it would be interesting to see where the query "craigslist" would be ranked in the list given the probability of this query will change in the US. Same applies for queries "ny lottery" and "Florida lottery" for the states of NY and FL respectively.

Identifying and removing any queries resulting from a search engine related feature or due to marketing: There may be some queries that may be there because of some search engine feature or due to a marketing related activity in a geographical location. These queries will skew the query distribution unnaturally. We need to identify and filter these queries out.

Query Preprocessing: More structured approach to taking query n-grams, normalization and stemming are also needed. Currently, the processing occasionally involves manual intervention to remove noise from the queries. Taking all n-grams with stemming and lemmatization and then computing log likelihood ratio for all n-grams would give us a better insight into the query distribution.

Query Thresholding: Currently we filter out the queries occurring less than N times from the input data set. We should also be able to exclude queries that are way too frequent. For instance queries like google or Wikipedia are navigational queries and we do not get much information except that people are looking for these sites. We should be able to optionally remove these queries from analysis.

Removing noise from the data: There are queries that clearly are noise. For instance, Hawaii has a query 'define <word>' that is one of the top queries biased toward Hawaii. We do not understand what the intention of the user here is? Also, we cannot explain why this query is showing up in top queries for Hawaii. The only explanation is that our bot filtering did not work correctly and let some of the bots pass through.

Using Bayesian Rule for Computing the Probabilities: We would like to explore other techniques such as simple Bayes rule to compute the probabilities. We realized that the probabilities could also be computed as follows

$$P(state = s|Q_i) = \frac{P(Q_i|state = s)\,P(state = s)}{P(Q_i)}$$

Alaska (AL)	Arizona(AZ)	California(CA)	Florida(FL)	Hawaii(HI)	Masachussetes(MA)	Michigan(MI)	New York(NY)	Texas(TX)	Washington(WA)
alaska airlines	azcentral	california lottery	florida lottery	friendster	craigslist boston	michigan lottery	ny lottery	mitchell's gourmet foods union	seattle traffic
define <*word*>	cox webmail	microsoft messenger	orkut	ako	boston	detroit news	newsday	texas lottery	craigslist seattle
levi johnston on tyra	cox	wamu	search live	mypay	citizens bank	nwa	optonline	houston chronicle	seattle times
eminem on palin	chase bank	zillow	suntrust	united airlines	health fitness	charter	barack obama	funbrain	chevrolet
susan boyle makeover	us airways	costco	bellsouth	usaa	buy xbox 360	windows live hotmail	jetblue	american airlines	msft
bristol palin custody	wells fargo	wells fargo	windows live hotmail	orbitz	tom tom gps	webkinz	citibank	univision	seattle weather
ako	chase	univision	music	define <*word*>	zune	chase	hi5	wells fargo	barack obama
usaa	southwest airlines	photobucket	wachovia	costco	windows live	comcast	meebo	chase	alaska airlines
wells fargo	costco	tmz	msnbc media	netflix	orkut	kohls	aim	att	buy xbox 360
sarah palin newt gingrich keynote republican dinner	southwest	club penguin	msn video	travelocity	verizon email	pogo	chase	disney channel	tom tom gps

Figure 3: Top 10 Most likely to occur search engine queries for the states of AL, AZ,CA,FL,HI,MA,MI,NY,TX,WA

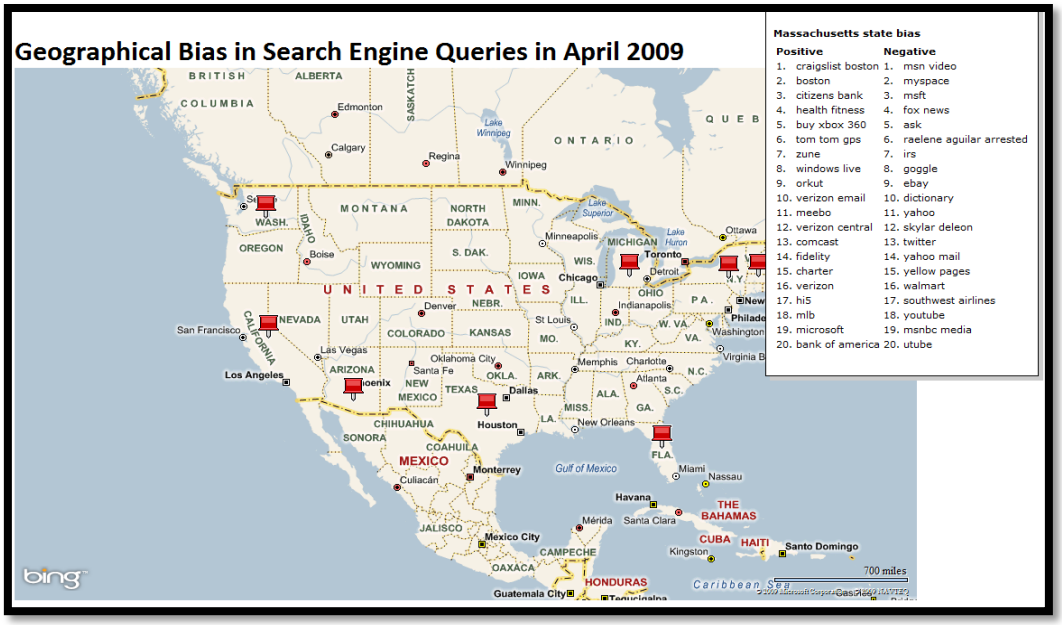

Figure 4: Screenshot from the demo showing the states in mainland US analyzed during the study. Positive and Negative columns in the fly out text show the queries that are most and least likely to occur in the state of MA respectively.

Mapping of Queries to Concepts: We should also have a way to collapse some queries into concepts such as facebook and myspace going into social networking and CNN and Fox going to news. This will help us understand the broader concept classes people are interested in. Currently, query term "Zillow" seems to indicate great interest in the housing market from the state of CA, which can be explained by the fact that many people may be interested in becoming first time homebuyers due to low housing prices. This information is helpful in helpful advertisers target their ads for better user engagement.

Zooming in and out of a region: Collapsing queries like 'San Jose traffic' and '95050 traffic' into one query, if we are analyzing on state level but keep them separate if analyzing on city level can give us better understanding of more granular query behavior.

Predict what is about to happen: Do we have the ability to predict happenings by analyzing the distribution of queries. If people are issuing a lot of queries related to flu related symptoms, should we expect to see a surge in flu like illnesses from the geographical region?

Arranging the terms in query lexicographically: For instance "barack obama nobel prize" would be the same as "nobel prize barack obama" i.e. both will be mapped to "barack obama nobel prize" (note the that barack, obama, nobel, prize is the order you would see in the dictionary).

Taking all N-grams of query term into consideration: For the query "barack obama nobel prize" we would consider all n grams, i.e. "barack", "barack obama", "barak obama nobel" and "barack obama nobel prize". This would give a richer query set and would account for the fact that people may issue the same query with each term in different order.

References

[1] A. Ng. Advice on applying machine learning algorithms. Available online at http://www.stanford.edu/class/cs229/materials/ML-advice.pdf

[2] G. Buehrer, J.Stokes, and K Chellapilla. A Large-scale Study of Automated Web Search Traffic**.** Proceedings of the 4th international workshop on Adversarial information retrieval on the web 2008

[3] Yan et. al. How much can Behavioral Targeting Help Online Advertising? WWW2009 MADRID