Google Correlate Whitepaper

Matt Mohebbi, Dan Vanderkam, Julia Kodysh,

Rob Schonberger, Hyunyoung Choi & Sanjiv Kumar

Draft Date: June 9, 2011

Trends in online web search query data have been shown useful in providing models of real world phenomena. However, many of these results rely on the careful choice of queries that prior knowledge suggests should correspond with the phenomenon. Here, we present an online, automated method for query selection that does not require such prior knowledge. Instead, given a temporal or spatial pattern of interest, we determine which queries best mimic the data. These search queries can then serve to build an estimate of the true value of the phenomenon. We present the application of this method to produce accurate models of influenza activity and home refinance rate in the United States. We additionally show that spatial patterns in real world activity and temporal patterns in web search query activity can both surface interesting and useful correlations.

Background

Web search activity has previously been shown useful for providing estimates of real-world activity in a variety of contexts, with the most common being health and economics. Examples in health include influenza[1,2,3,4,5,6,9], acute diarrhea[6], chickenpox[6], listeria[7], and salmonella[8]. Examples in economics include movie box office sales[9], computer game sales[9], music billboard ranking[9], general retail sales[10], automotive sales[10], home sales[10], travel[10], investor attention[11], and initial claims for unemployment[12].

Modeling real-world activity using web search data can provide a number of benefits. First, it can be more timely, especially when the alternative is not electronically collected. Influenza surveillance from the United States Centers for Disease Control and Prevention (CDC), Influenza Sentinel Provider Surveillance Network (ILINet) has a delay of one to two weeks[1]. For economic indicators like unemployment, this delay is measured in months[10]. In contrast, search data can "predict the present" since it is available as the target activity happens[10]. Second, query data has good temporal and spatial resolution. If an indicator of interest is incomplete (missing time periods or regions, coarser temporal or spatial resolution, etc.), query data can sometimes be used to fill in the gaps. For example, influenza rate data from ILINet is only published by the CDC at the national and regional level and is not published for the off season[13], but models based on query data can be used to provide estimates year-round and at a state and sometimes even city level, provided there is sufficient search activity at that level[1,14,15]. Third, there can be considerable expenses incurred in collecting data for traditional indicators. Finally, while Internet users do not represent a random sample of the United States population, this population has become increasingly less biased over time and now represents 77% of the adult population[16]. In the 18-29 subgroup, this number is almost 90%. This is in contrast to traditional landline phone surveys which must either under-represent this age group or blend in cell-phone survey data at considerable difficulty and expense[17].

Three Google tools have been released previously to enable access to aggregated online web search query data. Google Trends and Google Insights for Search are both real-time systems which provide temporal and spatial activity for a given query. However, they are both unable to automatically surface queries which correspond with a particular pattern of activity. Google Flu Trends provides estimates of Influenza-like Illness (ILI) activity in the United States, using models based on query data. These queries are selected from millions of possible candidates through an automated process[1]. Due to the computational requirements of this process, a batch-based distributed computing framework[18] was employed to distribute the task across hundreds of machines.

Google Correlate builds on this previous work. Google Correlate is a generalization of Flu Trends that allows for automated query selection across millions of candidate queries for any temporal or spatial pattern of interest. Similar to Trends and Insights for Search, Google Correlate is an online system and can surface its results in real time.

Data Summary

Using anonymized logs of Google web search queries submitted from January 2003 to present, we computed two different databases for Google Correlate:

us-weekly: temporal only: weekly time series data for the United States at a national level.

us-states: spatial only: state-by-state series data for the United States summed across all time.

Each database contains tens of millions of queries. For additional details, please see the Data section below.

Methods Summary

The objective of Google Correlate is to surface the queries in the database whose spatial or temporal pattern is most highly correlated (R^2) with a target pattern. Google Correlate employs a novel approximate nearest neighbor (ANN) algorithm over millions of candidate queries in an online search tree to produce results similar to the batch-based approach employed by Google Flu Trends but in a fraction of a second. For additional details, please see the Methods section below.

Flu Trends

Google Flu Trends produces estimates of ILI activity in the United States using query data. The Flu Trends modeling process is composed of two steps: variable selection and model building. Google Correlate can perform the variable selection and provide the associated time series data as a CSV download to enable the construction of a model using the selected queries. In this section we provide a test of the quality and computational power of Google Correlate, demonstrating that this automated system can be used to build a new Flu Trends model for the United States with comparable performance, but in a fraction of the time used to build the original Flu Trends model.

The baseline for this comparison the original regional Google Flu Trends model[1]. For these models, query selection was performed on the regional level, and a single set of queries was chosen to optimize the results across all regions. The values of the query time series were summed into a single input variable per region, and a model was fitted from the data across all nine regions. This model was built using weekly training data between 9/28/2003 and 3/11/2007 inclusive, and evaluated by computing the correlation between the resulting predictive estimates and the corresponding regional weekly truth data over the holdout period between 3/18/2007

to 5/11/2008.

While we sought to make a close comparison between the results of the Google Flu Trends methodology and modeling of ILI activity using Google Correlate, there are several differences between the methods employed. First, we worked with a different resolution for query selection. Since Google Correlate provides only national query time series data, we can only perform query selection on the national level. After the national-level query selection, we sum the query time series into a single explanatory variable and fit a linear model to the nine census regions. Second, we used a different cross-validation technique for variable selection in Google Correlate from the one used in Flu Trends.

We used Google Correlate to perform query selection by uploading ILI activity data from the CDC over the training time period. This weekly time series is at the national level and represents the rate of ILI-related doctors office visits per 100,000 visits. We summed the time series of all 100 queries returned by Google Correlate into a single explanatory variable. We then fit a linear model to the nine census regions and generated regional estimates for the holdout time period.

Training window correlation (R^2)

	Mean	*Min*	*Max*
Google Flu Trends	*0.90*	*0.80*	*0.96*
Google Correlate	*0.87*	*0.70*	*0.97*

n = 9 regions

Holdout window correlation (R^2)

	Mean	*Min*	*Max*
Google Flu Trends	*0.97*	*0.92*	*0.99*
Google Correlate	*0.96*	*0.88*	*0.98*

n = 9 regions

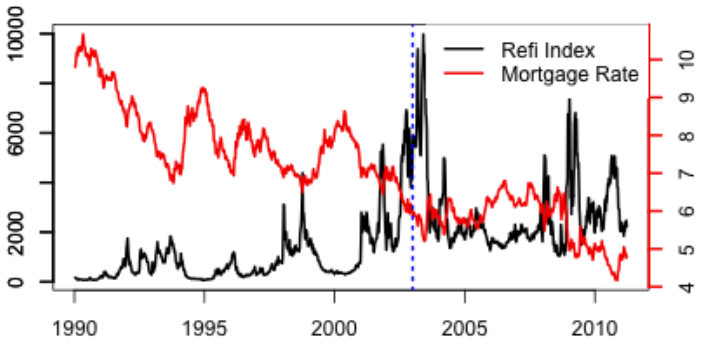
We see that the Google Correlate-based model slightly underperforms the Flu Trends model for the hold out time, with average correlation across all nine regions of 0.97 for Flu Trends and 0.96 for Correlate. This difference could be due, in part, to the difference in resolution of the query selection process. The time required to create the model with Google Correlate was a fraction of that required for the original Flu Trends model.

Refinance

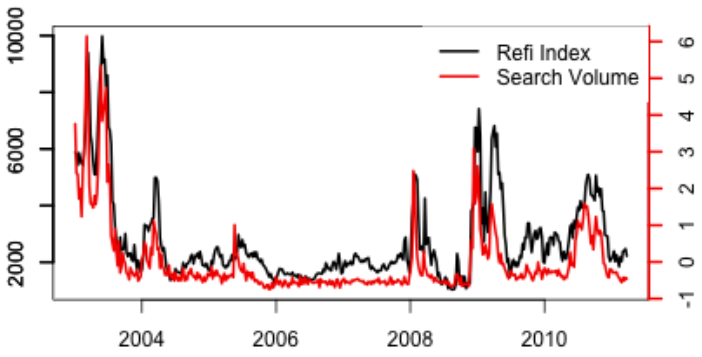
Every week, Mortgage Bankers Association of America (MBA) compiles all mortgage application to refinance an existing mortgage into a refinance index. The MBA's loan application survey covers more than half of all United States residential mortgage loan applications and is considered by many to be the best gauge of mortgage refinancing activity.

Consumers refinance a home for a number of reasons, including to switch to a lower mortgage interest rate, to change the mortgage length, to tap into their home equity and to switch mortgage type. In 2003, the refinancing activity peaked due to record low interest rate and the real estate boom. Despite the lower mortgage interest rate in 2010, the level of refinancing was not as high as in 2003 due to the housing recession and the subprime credit crisis.

We examined the top 100 most correlated queries with the refinance index time series from January 2003 to August 2010 and extended the window week by week until the end of March 2011. Fifty percent of the selected queries were refinance-related, including *refinancing calculator*, *refinancing closing costs*, and *refinance comparison*. Mortgage rate related queries such as *lowest mortgage rates* and *no cost mortgage* accounted for about 35% of queries selected. Even though queries for mortgage rates are related to refinancing, it is not always about refinancing and thus the signal could be mixed.



Refi Index vs. Mortgage Rate



Refi Index vs. Search Volume of *refinancing calculator*

Using these queries, we applied the same method from Choi and Varian[10] and compared two alternative models with baseline model with a moving window from August 2010 to March 2011. Let y_t be the time series of the refinance index, Refi_t be the summed query time series for queries returned by Google Correlate containing "refinance" or "refinancing", and Finance_t be the summed query time series for all 100 queries returned by Google Correlate.

Baseline Model: $y_t = \alpha + \phi y_{t-1} + e_t$
Alternative Model 1: $y_t = \alpha + \phi y_{t-1} + \beta \text{Refi}_t + e_t$
Alternative Model 2: $y_t = \alpha + \phi y_{t-1} + \beta \text{Finance}_t + e_t$

The model fit is significantly improved and prediction error is decreased for the two alternatives. The out of sample mean absolute error (MAE) with rolling window for the 31 weeks is decreased by 7.04% for Alternative Model 1 and the MAE for Alternative Model 2 is increased by 9.12%.

		ϕ	β	R^2	MAE
Baseline Model	Est	0.945		0.9011	8.42
	s.e.	0.015			
Alternative Model 1	Est	0.473	16.7	0.9363	7.61
	s.e.	0.033	1.1		
Alternative Model 2	Est	0.481	9.1	0.9366	7.82
	s.e.	0.032	0.6		

Ribosome

A ribosome is a component inside living cells. Using the *us-weekly* database, the query *ribosome* surfaces the following highly-correlated ($R^2 > 0.96$) queries:

1. *mitochondria*
2. *cell wall*
3. *chloroplasts*
4. *chromatin*
5. *plant cells*
6. *vacuole*
7. *chloroplast*
8. *nuclear membrane*
9. *reticulum*
10. *cell function*

The time series for these queries feature upticks in the Fall and Spring, sharp drops during Thanksgiving and Christmas and a long trough in the summer. This mirrors the school year in the United States and suggests that the queries are being driven by biology classes.

It is worth noting that all of these top terms relate to biology. Other school topics (e.g. the Canterbury Tales) are also studied early in the school semester and yet this time series is not correlate nearly as well. It's both surprising and impressive that the phenomenon of biology study appears to be uniquely characterized by its temporal pattern. This can be seen with other queries, for example *eigenvector*, but to a smaller extent.

Latitude

Using a *us-states* data series containing the latitude for each state in the United States, we find the following highly-correlated queries were surfaced ($R^2 > 0.84$):

1. *sad light therapy*
2. *defroster*
3. *seasonal affective disorder lights*
4. *10000 lux*
5. *sun lamp*
6. *track length*
7. *floor heating*
8. *fleece hat*
9. *irish water spaniel*
10. *hydronic*

The "sad" in *sad light therapy* is likely the acronym for seasonal affective disorder, which also seems to describe the relationship between queries *sad light therapy*, *seasonal affective disorder lights*, *10000 lux* and *sun lamp*. These top results surfaced by Google Correlate imply that latitude in the United States can be modeled using the spatial patterns in SAD-related queries. This is consistent with studies on the correlation of SAD prevalence and latitude in North America[19].

Disclaimers

This system is not intended to serve as a replacement for traditional data collection mechanisms. While the queries selected by Google Correlate for a specific target series exhibit strong correlations with the target series over many years, this correspondence may not hold in the future due to changes in user behavior which are unrelated to the target behavior. For example, the correlation of a drug whose time series historically tracked well the activity of a disease, could significantly be changed by a recall of the drug.

Additionally, the underlying cause of search behavior can never be known. Users submitting influenza-like illness (ILI) queries are not necessarily experiencing ILI-symptoms. And similarly, non-ILI related queries which are highly correlated with an ILI series do not necessarily increase or decrease the likelihood of contracting influenza.

Query data does not represent a random sample of the population. While over three quarters of United States adults use the Internet, several subgroups are underrepresented. This could lead to sampling error depending on the modeling performed.

Google Correlate requires indicators with unique spatial or temporal patterns. Indicators with little variation or with very regular variation are unlikely to surface meaningful results. Indicators with unique variation may still not surface results due to a lack of information-seeking behavior for the indicator.

Acknowledgements

The authors would like to thank Doug Beeferman and Jeremy Ginsberg for providing early inspiration for Google Correlate. We'd also like to thank Hal Varian for his valuable feedback on Google Correlate and Jean-Baptiste Michel for his useful comments on this manuscript. Finally, we'd like to thank Craig

Nevill-Manning and Corinna Cortes for their guidance and support.

Privacy

At Google, we recognize that privacy is important. None of the data in Google Correlate can be associated with a particular individual. The data contains no information about the identity, IP address, or specific physical location of any user. Furthermore, any original web search logs older than nine months are anonymized in accordance with Google's Privacy Policy[20].

Data

Google Correlate contains two different databases of Google web search queries. The first contains contains weekly time series for the United States at a national resolution (*us-weekly*). The second contains state-by-state series for the United States summed across all time (*us-states*). Both datasets are one-dimensional, with *us-weekly* having a time dimension but no space dimension and *us-states* having a space dimension but no time dimension. Both dataset contain tens of millions of series.

To help smooth query data across similar underlying user behavior, n-grams of the queries are used as series identifiers. This approach is similar to Google Trends and Insights for Search but is in contrast to Flu Trends where only lowercasing was performed on the queries.

The following example illustrates how n-grams are extracted from the query 'cold and flu symptoms'.

cold *
cold and
cold and flu *
cold and flu symptoms *
and *
and flu
and flu symptoms
flu
flu symptoms *
symptoms *

This list is filtered to contain only n-grams which appear often and in many states. The n-grams marked with an asterisk are kept when this filter is applied using the *us-weekly* dataset. Each of these filtered n-grams has a corresponding time series stored in the database, and for each instance of 'cold and flu symptoms' in the web search logs, each resulting n-gram receives a count. Filtering is done for privacy reasons but since rare queries are sporadic in nature, they are unlikely to be useful for modeling of long term phenomena. Distracting queries such as misspellings and those containing adult sexual content are also excluded.

The series in both datasets are normalized by dividing by the total count for all queries in that week (us-weekly) or state (*us-states*). The normalization controls for the year over year growth in all Internet search use (us-weekly) and state-by-state variation in Internet usage (*us-states*). Finally, each time series is standardized to have a mean value of zero and a variance of one, so that queries can be easily compared.

Methods

In our Approximate Nearest Neighbor (ANN) system, we achieve a good balance of precision and speed by using a two-pass hash-based system. In the first pass, we compute an approximate distance from the target series to a hash of each series in our database. In the second pass, we compute the exact distance function on the top results returned from the first pass.

Each query is described as a series in a high-dimensional space. For instance, for us-weekly, we use normalized weekly counts from January 2003 to present to represent each query in a 400+ dimensional space. For *us-states*, each query is represented as a 51-dimensional vector (50 states and the District of Columbia). Since the number of queries in the database is in the tens of millions, computing the exact correlation between the target series and each database series is costly. To make search feasible at a large scale, we employ an ANN system that allows fast and efficient search in high-dimensional spaces.

Traditional tree-based nearest neighbors search methods are not appropriate for Google Correlate due to the high dimensionality which results in sparsenes. Most of these methods reduce to brute force linear search with such data. For Google Correlate, we used a novel asymmetric hashing technique which uses the concept of projected quantization[21] to reduce the search complexity. The core idea behind projected quantization is to exploit the clustered nature of the data, typically observed with various real-world applications. At the training time, the database query series are projected in to a set of lower dimensional spaces.

Each set of projections is further quantized using a clustering method such as K-means. K-means is appropriate when the distance between two series is given by Euclidean distance. Since Pearson correlation can be easily converted into Euclidean distance by normalizing each series to be a standard Gaussian (mean of zero, variance of one) followed by a simple scaling (for details, see appendix), K-means clustering gives good quantization performance with the Google Correlate data. Next, each series in the database is represented by the center of the corresponding cluster.

This gives a very compact representation of the query series. For instance, if 256 clusters are generated, each query series can be represented via a unique ID from 0 to 255. This requires only 8 bits to represent a vector. This process is repeated for each set of projections. In the above example, if there are m sets of projections, it yields an 8m bit representation for each vector.

During the online search, given the target series, the most correlated database series are retrieved by asymmetric matching. The key concept in asymmetric matching is that the target query is not quantized but kept as the original series. It is compared against the quantized version of each database series. For instance, in our example, each database series is represented as an 8m bit code. While matching, this code is expanded by replacing each of the 8 bits by the corresponding K-means center obtained at training time, and Euclidean distance is computed between the target series and the expanded database series. The sum of the Euclidean distances between the target series and the database series in m subspaces represents the approximate distance between the two. Approximate distance between target series and the database series is used to rank all the database series. Since the number of centers is usually small, matching of the target series against all the database series can be done very quickly.

To further improve the precision, we take the top one thousand series from the database returned by our approximate search system (the first pass) and reorder those by doing exact correlation computation (the second pass). By combining asymmetric hashes and reordering, the system is able to achieve more than 99% precision for the top result at about 100 requests per second on O(100) machines, which is orders of magnitude faster than exact search.

References

1. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. (2009) Detecting influenza epidemics using search engine query data. Nature 457: 1012-1014.

2. Eysenbach G (2006) Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA Annu Symp Proc: 244-248.

3. Hulth A, Rydevik G, Linde A (2009) Web queries as a source for syndromic surveillance. PLoS One 4: e4378-e4378.

4. Johnson HA, Wagner MM, Hogan WR, Chapman W, Olszewski RT, et al. (2004) Analysis of Web access logs for surveillance of influenza. Stud Health Technol Inform 107: 1202-1206.

5. Polgreen PM, Chen Y, Pennock DM, Nelson FD (2008) Using internet searches for influenza surveillance. Clin Infect Dis 47: 1443-1448.

6. Pelat C, Turbelin Cm, Bar-Hen A, Flahault A, Valleron A-J (2009) More diseases tracked by using Google Trends. Emerg Infect Dis 15: 1327-1328.

7. http://ecmaj.ca/cgi/content/full/180/8/829

8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2917042/

9. http://www.pnas.org/content/107/41/17486.full.pdf

10. http://www.google.com/googleblogs/pdfs/google_predicting_the_present.pdf

11. http://www.nd.edu/~zda/Google.pdf

12. http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/archive/papers/initialclaimsUS.pdf

13. http://www.cdc.gov/flu/weekly

14. http://www.eht-journal.net/index.php/ehtj/article/view/7183/8094

15. http://www.nature.com/nature/journal/v457/n7232/extref/nature07634-s1.pdf

16. http://www.pewinternet.org/Static-Pages/Trend-Data/Whos-Online.aspx

17. http://pewresearch.org/pubs/515/polling-cell-only-problem

18. Dean, J. & Ghemawat, S. Mapreduce: Simplified data processing on large clusters. OSDI: Sixth Symposium on Operating System Design and Implementation (2004)

19. http://cbn.eldoc.ub.rug.nl/FILES/root/1999/JAffectDisordMersch/1999JAffectDisordMersch.pdf

20. http://www.google.com/privacypolicy.html

21. A. Gersho and R. M. Gray, Vector Quantization and Signal Compression, Springer, 1991.