Tools for Google Data

Hal Varian
Google
July 2015
Google Trends

Google Correlate

Google Consumer Surveys
Which day of the week are there the most searches for [hangover]?

1: Sunday
2: Monday
3: Tuesday
4: Wednesday
5: Thursday
6: Friday
7: Saturday
Search index for [hangover]


Interest over time
The number 100 represents the peak search volume
Hangover-vodka time series


Interest over time

The number 100 represents the peak search volume

Average

Dec 2008

Jan 2009
Granger Causality?

Took a 7-day difference to eliminate seasonal component and used lag = 1 (based on auto.arima).

Does vodka Granger-cause hangovers? Yes, p=2.7e-13
Do hangovers Granger-cause vodka? No, p=0.4588
Searches for [civil war]


Interest over time

The number 100 represents the peak search volume

Note
Searches for [term paper]


Interest over time
The number 100 represents the peak search volume

Note

Embed
Gift for boyfriend v Gift for girlfriend

Interest over time
The number 100 represents the peak search volume

For boyfriend 
For girlfriend
"Calling the Greek referendum on the Nose with Google Trends", by Nikos Askitas, available at SSRN. Used *hourly* Trends data comparing “yes” and “no” votes.

### CALLING THE GREFERENDUM

**TABLE I**

**Top searches for NO** i.e. “οχι - ναι”

<table>
<thead>
<tr>
<th>Greek</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>βεγγος οχι</td>
<td>Veggos No</td>
</tr>
<tr>
<td>δημοψηφισμα οχι</td>
<td>referendum no</td>
</tr>
<tr>
<td>οχι στο δημοψηφισμα</td>
<td>no in the referendum</td>
</tr>
<tr>
<td>συγκεντρωση υπερ του οχι</td>
<td>demonstration in favour of no</td>
</tr>
<tr>
<td>υπερ του οχι</td>
<td>in favour of no</td>
</tr>
<tr>
<td>συγκεντρωση ναι</td>
<td>demonstration no</td>
</tr>
<tr>
<td>λεμε οχι</td>
<td>we say no</td>
</tr>
<tr>
<td>δημοψηφισμα οχι</td>
<td>referendum no</td>
</tr>
<tr>
<td>οχι</td>
<td>no</td>
</tr>
<tr>
<td>ψηφιζω οχι</td>
<td>I vote no</td>
</tr>
</tbody>
</table>

**TABLE II**

**Top searches for YES** i.e. “ναι - οχι”

<table>
<thead>
<tr>
<th>Greek</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>ναι στην Ευρωπη</td>
<td>yes to Europe</td>
</tr>
<tr>
<td>δημοψηφισμα ναι</td>
<td>referendum yes</td>
</tr>
<tr>
<td>ναι στο δημοψηφισμα</td>
<td>yes in the referendum</td>
</tr>
<tr>
<td>συγκεντρωση υπερ του ναι</td>
<td>demonstration in favour of yes</td>
</tr>
<tr>
<td>συγκεντρωση για το ναι</td>
<td>demonstration for yes</td>
</tr>
<tr>
<td>ναι στο ευρω</td>
<td>yes to euro</td>
</tr>
<tr>
<td>επιτροπη υποστηριξης του ναι</td>
<td>committee in support of yes</td>
</tr>
</tbody>
</table>
Google Searches on “yes” and “no” votes (hourly)

Source: Askitas [2015]
Ratio of “no” to “yes”

Predicted share: 60%     Actual share: 61.31%

Source: Askitas [2015]
Google Trends
Google Correlate
Google Consumer Surveys
Searches correlated with [weight loss]

Correlated with **weight loss**
- 0.9603 **loss**
- 0.9270 **weight**
- 0.8851 **losing weight**
- 0.8706 **best vacation spots**
- 0.8705 **low calorie**
- 0.8580 **condos for**
- 0.8578 **best resorts**
- 0.8568 **weight loss pills**
- 0.8556 **best vacation**
- 0.8555 **body fat percentage**
Plot of [weight loss] and [best vacation spots]

United States Web Search activity for **weight loss** and **best vacation spots** ($r=0.8706$)

*Hint: Drag to zoom, and then correlate over that time only.*

**New Year**
Correlated with [weight loss] 3 weeks later
What can you do with Trends and Correlate?

**Time series nowcasting**
- Forecast unemployment insurance claims
- Forecast/nowcast home sales in the US

**Cross section correlations**
- Model cities with largest drop in housing prices
- What kinds of queries are associated with happy cities?
- What kinds of queries are associated with places with short lifespans?
- Where does “assembled in America” resonate?
Initial claims: good leading indicator for recessions

Grey bars indicate recessions
Google Correlate with initial claims data

- Compare weekly time series
- Compare monthly time series

Shift series: 0 weeks
Country: United States

Correlated with Initial claims NSA:
- 0.8679 michigan unemployment
- 0.8273 idaho unemployment
- 0.8222 pennsylvania unemployment
- 0.8114 unemployment filing
- 0.8061 new jersey unemployment
- 0.8020 illinois unemployment
- 0.8017 department of unemployment
- 0.8012 rhode island unemployment
- 0.7939 unemployment office
- 0.7933 filing unemployment

Documentation:
- Comic Book
- FAQ
- Tutorial
- Whitepaper

Show more | Export data as CSV | Share: Google+ Twitter
Initial claims and [unemployment filing]

User uploaded activity for Initial claims NSA and United States Web Search activity for unemployment filing
(r=0.8114)

Hint: Drag to Zoom, and then correlate over that time only.
Nowcasting initial claims

Predict NSA initial claims \( (y_t) \), using lagged values of initial claims and contemporaneous queries on [unemployment filing] \( (x_t) \)

Base: \[ y_t = a_0 + a_1 y_{t-1} + a_{52} y_{t-52} + e_t \]

Trends: \[ y_t = a_0 + a_1 y_{t-1} + a_{52} y_{t-52} + b x_t + e_t \]

Result: R\(^2\) goes from 80.8% to 87.6%

Use `auto.arima` from Ron Hyndman's `forecast` package. Result: model essentially the same with and without \( x_t \) and AIC goes from 191.53 to 177.08
How can we make variable selection easier?

How to choose best predictors?
- Simple correlation?
- Judgment?
- Stepwise regression?
- Lasso, LARS, Elastic Net?

Spike-and-slab regression
- Kalman filter for trend and seasonality
  - Prior probability variable is included (spike)
  - Prior probability distribution over coefficient value (slab)
- Sample from simulated posterior, average to get prediction
- See Scott and Varian (2012, 2013) for details
- Download R package from CRAN (BoomSpikeSlab, bsts)
New Home Sales in US

Correlated with HSN1FNSA
- tahitian noni juice (0.9821)
- exhaust sound (0.9808)
- traderonline.com (0.9800)
- www.kbb.com (0.9791)
- 80/20 mortgage (0.9790)
- appreciation rate (0.9786)
- home appreciation (0.9780)
- help-u-sell (0.9764)
- new home builder (0.9762)
- bostonworks.com (0.9762)

User uploaded activity for HSN1FNSA and United States Web Search activity for 80/20 mortgage (r=0.9790)
User uploaded activity for HSN1FNSA and United States Web Search activity for 80/20 mortgage ($r=0.9790$)

Hint: Drag to Zoom, and then correlate over that time only.
Predictors chosen by model

- appreciation.rate
- irs.1031
- century.21.realtors
- real.estate.purchase
- X80.20.mortgage
- selling.real.estate
- estate.appraisal
- real.estate.appraisal

Inclusion Probability
Incremental fit plots

Visualize how much each predictor contributes to model fit

model: $y_t = \text{trend}_t + \text{seasonal}_t + b_1 x_{1t} + b_2 x_{2t}$

plot1: $y_t = \text{trend}_t$

plot2: $y_t = \text{trend}_t + \text{seasonal}_t$

plot3: $y_t = \text{trend}_t + \text{seasonal}_t + b_1 x_{1t}$

plot4: $y_t = \text{trend}_t + \text{seasonal}_t + b_1 x_{1t} + b_2 x_{2t}$
1. trend (mae=0.51964)
2. add seasonal (mae=0.5173)
3. add appreciation.rate (mae=0.24611)
4. add irs.1031 (mae=0.1635)
5. add century.21.realtors (mae=0.1153)
6. add real.estate.purchase (mae=0.087582)
7. add X80.20.mortgage (mae=0.064159)
One month ahead forecast

Does 23% better than simple AR1 model
You can also do cross section prediction

Upload data on house price declines into Google Correlate.

Highest correlation is queries on [short sale process]
Find best predictors using spike-and-slab
Compare actual to fitted
Where are the happiest cities in the US?

Ed Glaeser, Joshua Gottlieb, and Oren Ziv, “Unhappy Cities”, NBER 20291

CDC Behavioral Risk Factor Surveillance System survey: “In general, how satisfied are you with your life?” Answers for 174 cities.

What kinds of queries are predictive of happy and unhappy cities?
Probability of inclusion in regression for happy cities

Inclusion Probability

- Gambling
- Manufacturing
- Insurance
- Coupons...Discount.Offers
- Toys
- Home.Furnishings
- Health.Insurance
- MalePopulation
- Card.Games
- Boats...Watercraft
Scatterplot of actual and predicted
Scatterplot with labels
Residents of Fairfax County are among the longest-lived in the country: Men have an average life expectancy of 82 years and women, 85, about the same as in Sweden. In McDowell [County], the averages are 64 and 73, about the same as in Iraq.”

Question: what queries are predictive of morbidity at state level? Use Google Correlate!
State-level correlations

User uploaded activity for negative life expectancy and United States Web Search activity for blood pressure medicine ($r=0.9092$)
Predictors of morbidity

- he is god
- there are 57 states
- the Constitution is dead
- serve satan
- uh too much
Google Trends
Google Correlate
Google Consumer Surveys
Consumer Surveys

1. You create online surveys to gain consumer insight.
2. People complete questions to access premium content.
3. Publishers get paid as their visitors answer.
4. You get nicely aggregated and analyzed data.
Data Mining: The Big Dig

Posted on June 11, 2000

Frontier: Instant Expert

Data Mining: The Big Dig

Your databases and Web sites hold vast stores of information on customer buying habits and market trends—if you know how to analyze the patterns. Some entrepreneurs are intimidated by technical issues or price: Hiring a pro for sophis...

Answer a question to continue reading this page

question 2 of 2:
Have you ever purchased anything from (check all that apply):

Choose all answers that apply

☐ An email newsletter or ad
☐ A YouTube video
☐ An ad on your mobile phone
☐ An ad on your tablet
☐ None of the above

Submit answer(s)

Show me another question
How this changes surveys

Anyone can do them

The cost is dramatically lower

Results come back in a few hours

Surveys can be replicated … or not

You can measure sensitivity to wording
I prefer to buy products that are assembled in America

Results for respondents with demographics. Weighted by Age, Gender, Region. (2116 responses)

Confidence too close to call.

- Strongly agree: 32.9% (+2.1 / -2.1)
- Agree: 30.5% (+2.1 / -2.0)
- Neutral: 25.8% (+2.0 / -1.9)
- Mildly disagree: 3.2% (+1.0 / -0.8)
- Strongly disagree: 7.6% (+1.3 / -1.1)
Predictors of survey response
## Top and bottom cities' predicted score

<table>
<thead>
<tr>
<th>Top</th>
<th>Bottom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kershaw, SC: 83.2 %</td>
<td>Calipatria, CA: 40.2 %</td>
</tr>
<tr>
<td>Summersville, WV: 82.8 %</td>
<td>Fremont, CA: 40.2 %</td>
</tr>
<tr>
<td>Grundy, VA: 82.8 %</td>
<td>Mountain View, CA: 40.8 %</td>
</tr>
<tr>
<td>Chesnee, SC: 82.7 %</td>
<td>San Jose, CA: 41.4 %</td>
</tr>
<tr>
<td>Duffield, VA: 82.5 %</td>
<td>Berkeley, CA: 41.4 %</td>
</tr>
<tr>
<td>Norton, VA: 82.3 %</td>
<td>Redmond, WA: 41.5 %</td>
</tr>
<tr>
<td>Jonesville, VA: 82.2 %</td>
<td>Glendale, CA: 41.5 %</td>
</tr>
<tr>
<td>Walnut Cove, NC: 82.2 %</td>
<td>Cupertino, CA: 41.6 %</td>
</tr>
<tr>
<td>Weston, WV: 82.2 %</td>
<td>Palo Alto, CA: 41.7 %</td>
</tr>
<tr>
<td>Ennice, NC: 82.1 %</td>
<td>Daggett, CA: 41.9 %</td>
</tr>
</tbody>
</table>
Geo-amplification

You can do the same thing for any geographically distribute variable

Find out queries or query categories that are predictive of that variable

Make predictions/extrapolations to other geographies

Many applications

Social science

Policy

Marketing

Politics

Example: *New York Times index* of “hard places” (June 26, 2014)
Where are the hardest places to live in the U.S.?
What queries are associated with “hard places”? 

Based on state level data and Google Correlate
What queries are associated with “easy places”? Based on state level data and Google Correlate.
Challenges for the future

Private sector has high-frequency, real time data and a lot of it!

- Visa, Mastercard, American Express
- UPS and FedEx
- Wal-Mart, Target, etc
- Supermarket scanner data
- Search engines

Government agencies

- Long historical series, but usually low frequency
- Carefully constructed but labor intensive, with delayed release and periodic revisions

How to combine the public and private data?

- How to integrate massive amounts of private sector real-time information with traditional government statistics
Unused slides
One more nowcasting example: consumer sentiment
What query categories are the best predictors of consumer sentiment?
Probability of inclusion of predictor (n=98, k=195)

- Financial.Planning
- Investing
- Business.News
- SearchEngines
- EnergyUtilities

- White: procyclical
- Black: countercyclical
Start with “trend”
Add “financial planning”
Add “Investing”
Add “Business News”
Add “Search Engines”
Add Energy and Utilities
Predictors of morbidity