Forecasting and Google Flu Trends (02-05)

Matthew J. Salganik
Department of Sociology
Princeton University

Soc 596: Computational Social Science

prediction vs forecasting

Rain dance problems vs umbrella problems (Kleinberg et al 2015)

Fenn and Raskino (2008)

Watch the progression . . . . .

Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance

Gunther Eysenbach MD MPH [1) 2)]
1) Centre for Global eHealth Innovation, University Health Network, Toronto M5G2C4, Canada, and 2) Department of Health Policy, Management and Evaluation, University of Toronto, Canada.
Email: geysenba@uhnres.utoronto.ca

Eysenbach (2006) "Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance", American Medical Informatics Association Annual Symposium Proceedings,
`http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839505/`

The “trick” used here is that – while Google normally does not provide detailed log files for “all searches” conducted on its sites – it does provide rather detailed statistics for advertisers who “buy” (or rather bid for) certain keywords in the context of its keyword-triggered advertising program Google Adsense. (Note that keywords

To obtain statistics on the prevalence of searches on a certain topic I therefore created a "campaign" using a keyword-triggered "sponsored link" in Google Adsense, which appeared for Canadian searchers only, who entered "flu" or "flu symptoms" into Google. The ad read "Do you have the flu? Fever, Chest discomfort, Weakness, Aches, Headache, Cough." and contained a link to a generic patient education website.

Figure 1. Normalized data from Fluwatch (influenza cases, lab tests, ILI reports from sentinel physicians) and Google (number of clicks on an keyword-triggered influenza link).

Aside: Interesting use of ad-click rates.

Using Internet Searches for Influenza Surveillance

Philip M. Polgreen,[1,2] Yiling Chen,[4] David M. Pennock,[5] and Forrest D. Nelson[3]

[1]Department of Internal Medicine, Carver College of Medicine, [2]Department of Epidemiology, College of Public Health, and [3]Department of Economics, Henry B. Tippie College of Business, University of Iowa, Iowa City; [4]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts; and [5]Yahoo! Research, New York, New York

Polgreen et al. (2008) "Using Internet Searches for Influenza Surveillance", *Clinical Infectious Diseases*, `http://dx.doi.org/10.1086/593098`

We obtained 2 series of influenza-related search fraction data at the national level: (1) the fraction of US search queries that contain the terms “influenza” or “flu” but do not contain the terms “bird,” “avian,” or “pandemic” and (2) the fraction of US search queries that contain the terms “influenza” or “flu” but do not contain the terms “bird,” “avian,” “pandemic,” “vaccine,” “vaccination,” or “shot.”

Note that they had access to raw search data from Yahoo!

Influenza-related searches, %
0
.00005
.0001
.00015
Cultures positive for influenza, %
0
10
20
30
1–2005
1–2006
1–2007
1–2008
Week–Year
Internet searches
Cultures positive for influenza

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

Ginsberg et al. (2009) "Detecting influenza epidemics using search engine query data", *Nature*, `http://dx.doi.org/10.1038/nature07634`

Our proposed system builds on this earlier work by using an automated method of discovering influenza-related search queries.

Key idea from data science:
Over-fitting

Which line fits the data better?

1. straight line
2. curved line

https://commons.wikimedia.org/wiki/File:Overfitted_Data.png

Keyword selection method:

- Try 50 million keywords and see which ones work best

Keyword selection method (roughly):

- For each possible keyword
- For each region
- Split 128 outcome measures (CDC flu measurements) into 4 sets of 96 outcome measures
- Fit linear regression: $I(t) = \beta_0 + \beta_1 Q(t)$
- Use fitted weights to predict held out 32 values ($96 + 32 = 128$)
- Calculate correlation between predicted values and actual values
- Average (transformed) correlations across cross validations and regions

Then, find sets of keywords that work well together

Table 1 | Topics found in search queries which were found to be most correlated with CDC ILI data

Search query topic	Top 45 queries		Next 55 queries	
	n	Weighted	*n*	Weighted
Influenza complication	11	18.15	5	3.40
Cold/flu remedy	8	5.05	6	5.03
General influenza symptoms	5	2.60	1	0.07
Term for influenza	4	3.74	6	0.30
Specific influenza symptom	4	2.54	6	3.74
Symptoms of an influenza complication	4	2.21	2	0.92
Antibiotic medication	3	6.23	3	3.17
General influenza remedies	2	0.18	1	0.32
Symptoms of a related disease	2	1.66	2	0.77
Antiviral medication	1	0.39	1	0.74
Related disease	1	6.66	3	3.77
Unrelated to influenza	0	0.00	19	28.37
Total	45	49.40	55	50.60

The top 45 queries were used in our final model; the next 55 queries are presented for comparison purposes. The number of queries in each topic is indicated, as well as query-volume-weighted counts, reflecting the relative frequency of queries in each topic.

Key idea from data science:
In-sample testing vs out-of-sample testing

Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.

Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.

Figure 3 | ILI percentages estimated by our model (black) and provided by the CDC (red) in the mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season. During week 5 we detected a sharply increasing ILI percentage in the mid-Atlantic region; similarly, on 3 March our model indicated that the peak ILI percentage had been reached during week 8, with sharp declines in weeks 9 and 10. Both results were later confirmed by CDC ILI data.

Privacy. None of the queries in the Google database for this project can be associated with a particular individual. The database retains no information about the identity, internet protocol (IP) address, or specific physical location of any user. Furthermore, any original web search logs older than 9 months are being made anonymous in accordance with Google's privacy policy (http://www.google.com/privacypolicy.html).

Privacy. None of the queries in the Google database for this project can be associated with a particular individual. The database retains no information about the identity, internet protocol (IP) address, or specific physical location of any user. Furthermore, any original web search logs older than 9 months are being made anonymous in accordance with Google's privacy policy (http://www.google.com/privacypolicy.html).

- Make a deontological argument against doing this study without consent.

Privacy. None of the queries in the Google database for this project can be associated with a particular individual. The database retains no information about the identity, internet protocol (IP) address, or specific physical location of any user. Furthermore, any original web search logs older than 9 months are being made anonymous in accordance with Google's privacy policy (http://www.google.com/privacypolicy.html).

- Make a deontological argument against doing this study without consent.
- Make a consequentialist argument for doing this study without consent.

Which do you prefer?

1. Researcher selected keywords (Polgreen et al., 2008)
2. Data selected keywords (Ginsberg et al., 2009)

Researchers selected key-words (Polgreen et al., 2008)

- takes advantage of researcher knowledge
- limits risk of over-fitting
- can be done without enormous amounts of data

Researchers selected key-words (Polgreen et al., 2008)

- takes advantage of researcher knowledge
- limits risk of over-fitting
- can be done without enormous amounts of data

Data-driven keyword selection (Ginsberg et al., 2009)

- generalized to every time-series all over the world
- potentially discovers surprisingly good predictors (e.g., antibiotic related queries)
- increases risks over-fitting
- requires lots of data and computing

Computation and pre-filtering. In total, we fit 450 million different models to test each of the candidate queries. We used a distributed computing framework[12] to divide the work among hundreds of machines efficiently. The amount of computation required could have been reduced by making assumptions about which queries might be correlated with ILI. For example, we could have attempted to eliminate non-influenza-related queries before fitting any models. However, we were concerned that aggressive filtering might accidentally eliminate valuable data. Furthermore, if the highest-scoring queries seemed entirely unrelated to influenza, it would provide evidence that our query selection approach was invalid.

What do you think?

1. Game changer
2. Incremental improvement

Readymades

Custommades

https://commons.wikimedia.org/wiki/File:Duchamp_Fountaine.jpg
https://commons.wikimedia.org/wiki/File:%27David%27_by_Michelangelo_JBU0001.JPG

Predicting consumer behavior with Web search

Sharad Goel[1], Jake M. Hofman[1], Sébastien Lahaie[1], David M. Pennock[1], and Duncan J. Watts[1]

Microeconomics and Social Systems, Yahoo! Research, 111 West 40th Street, New York, NY 10018

Goel et al. 2010. "Predicting consumer behavior with web search." *PNAS*, `http://dx.doi.org/XXX`

$$\mathbf{flu}_t = \beta_0 + \beta_1 \mathbf{flu}_{t-2} + \beta_2 \mathbf{flu}_{t-3} + \epsilon.$$

Fig. 4. Actual and estimated flu levels in the United States, where flu level is the percentage of physician visits that involve patients with influenza-like illnesses. Search-based estimates are from Google Flu Trends.

Did this change how you think about Ginsberg et al 2009?

Can search predict revenue?

Can search predict revenue?

Can search predict revenue?

Wrong question

How much can search improve simple baseline prediction?

How much can search improve simple baseline prediction?

Fig. 3. Predictions from the baseline (*A–C*) and the combined baseline-plus-search models (*D–F*) for movies, video games, and music.

How much can search improve simple baseline prediction?

Fig. 3. Predictions from the baseline (*A–C*) and the combined baseline-plus-search models (*D–F*) for movies, video games, and music.

Right question

Fig. 5. The correlation between predicted and actual outcomes for movies, video game sequels and nonsequels, music, and flu.

Key idea from data science:
Predictions have to be compared to something else

Key idea from data science:
Predictions have to be compared to something else
Tamara: the app you mentioned

Aside: See how the figures tell the story of this paper.

Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic

Samantha Cook[1], Corrie Conrad[2]*, Ashley L. Fowlkes[3], Matthew H. Mohebbi[1]

1 Google, Inc., New York, New York, United States of America, **2** Google, Inc., London, United Kingdom, **3** Influenza Division, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America

Cook et al. (2011) "Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic." *PLoS ONE*, `http://dx.doi.org/10.1371/journal.pone.0023610`

B ILINet Data and GFT Estimates: 2003 - 2009
ILINet
Erratic Period
Updated GFT Estimate
Original GFT Estimate
Beginning of H1N1 Wave 1
Beginning of H1N1 Wave 2
ILI Rate
0.00
0.02
0.04
0.06
0.08
2003-09-28
2004-09-26
2005-09-25
2006-09-24
2007-09-23
2008-09-21
2009-09-20

A ILINet Data and GFT Estimates: 2009
ILINet
Erratic Period
Updated GFT Estimate
Original GFT Estimate
Beginning of H1N1 Wave 1
Beginning of H1N1 Wave 2
Updated training data cutoff
ILI Rate
0.00
0.02
0.04
0.06
0.08
2009-01-11
2009-03-29
2009-08-02

Query Category	Sample Query	Original Model Relative Category Volume	Updated Model Relative Category Volume
Symptoms of an influenza complication	[symptoms of bronchitis]	6%	11%
Influenza complication	[pnumonia]*	42%	6%
Specific influenza symptom	[fever]	6%	39%
General influenza symptoms	[early signs of the flu]	2%	30%
Cold/flu remedy	[robitussin]	12%	4%
Term for influenza	[influenza a]	<1%	3%
Antibiotic medication	[amoxicillin]	12%	0%
Related disease	[strep throat]	16%	<1%

*Search users often misspell the word *pneumonia*.

doi:10.1371/journal.pone.0023610.t001

Conclusions: Internet search behavior changed during pH1N1, particularly in the categories "influenza complications" and "term for influenza." The complications associated with pH1N1, the fact that pH1N1 began in the summer rather than winter, and changes in health-seeking behavior each may have played a part. Both GFT models performed well prior to and during pH1N1, although the updated model performed better during pH1N1, especially during the summer months.

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

Lazer et al (2014) "The Parable of Google Flu: Traps in Big Data Analysis." *Science*.
http://dx.doi.org/10.1126/science.1248506

Google Flu
Lagged CDC
Google Flu + CDC
CDC
Google estimates more than double CDC estimates
% ILI
10
8
6
4
2
0
07/01/09
07/01/10
07/01/11
07/01/12
07/01/13

Google Flu
Lagged CDC
Google Flu + CDC
Google starts estimating high 100 out of 108 weeks
Error (% baseline)
150
100
50
0
−50
07/01/09
07/01/10
07/01/11
07/01/12
07/01/13
Data

What are the differences between these plots?

- Big data hubris
  Basically not paying attention

- Big data hubris
  Basically not paying attention
- Algorithm dynamics (combines what I called algorithmic confounding and drift)
  Adding recommended searches to increase health related queries

Projecting from the past to the future only works if nothing in the world changes; there is no model, no mechanism for understanding; this is blind curve-fitting

- Transparency and replicability

- Transparency and replicability What do you think Google should do in this case?
- Using big data to understand the unknown (basically estimate quantities that can't be estimated otherwise)

- Transparency and replicability What do you think Google should do in this case?
- Using big data to understand the unknown (basically estimate quantities that can't be estimated otherwise)
- Study the algorithm (Malte and Arvind)

- Transparency and replicability What do you think Google should do in this case?
- Using big data to understand the unknown (basically estimate quantities that can't be estimated otherwise)
- Study the algorithm (Malte and Arvind)
- It's not just about size of the data (calls for a hybrid between "big" and "small" data researchers)

Aside: note how this more general framing improves impact of paper

Accurate estimation of influenza epidemics using Google search data via ARGO

Shihao Yang[a], Mauricio Santillana[b,c,1], and S. C. Kou[a,1]

Yang et al (2015) "Accurate estimation of influenza epidemics using Google search data via ARGO", *PNAS*, `http://dx.doi.org/10.1073/pnas.1515373112`

Table 1. Comparison of different models for the estimation of influenza epidemics

	Whole period (Mar 29, 2009 to Jul 11, 2015)	Off-season flu H1N1	Regular flu seasons (week 40 to week 20 next year)				
			2010–2011	2011–2012	2012–2013	2013–2014	2014–15
RMSE							
ARGO	**0.608**	0.640	**0.596**	**0.807**	**0.687**	**0.306**	**0.438**
GFT (Oct 2014)	2.216	0.773	1.110	3.023	4.451	0.986	0.700
Ref. 16	0.915	0.833	0.881	2.027	1.090	0.446	0.663
GFT+AR(3)	0.912	**0.580**	0.602	1.382	1.279	0.993	0.906
AR(3)	0.957	0.813	0.794	1.051	1.191	0.969	0.928
Naive	1 (0.348)	1 (0.600)	1 (0.339)	1 (0.163)	1 (0.499)	1 (0.350)	1 (0.465)
MAE							
ARGO	**0.649**	0.584	**0.574**	**0.748**	**0.650**	**0.391**	**0.530**
GFT (Oct 2014)	1.834	0.777	1.260	3.277	5.028	0.891	0.770
Ref. 16	1.052	0.719	1.010	2.211	1.029	0.610	0.820
GFT+AR(3)	0.888	**0.570**	0.613	1.308	1.016	1.034	0.839
AR(3)	0.925	0.777	0.787	0.951	0.988	0.917	0.934
Naive	1 (0.201)	1 (0.425)	1 (0.259)	1 (0.135)	1 (0.325)	1 (0.212)	1 (0.295)
MAPE							
ARGO	**0.787**	**0.620**	**0.663**	**0.770**	**0.719**	**0.453**	**0.620**
GFT (Oct 2014)	1.937	0.721	1.394	3.442	5.419	0.892	0.895
Ref. 16	1.381	0.765	1.380	2.306	1.251	0.754	0.958
GFT+AR(3)	1.037	0.683	0.698	1.407	0.986	1.062	0.828
AR(3)	1.003	0.894	0.814	0.947	0.939	0.891	0.916
Naive	1 (0.090)	1 (0.139)	1 (0.105)	1 (0.081)	1 (0.110)	1 (0.084)	1 (0.097)
Correlation							
ARGO	**0.986**	0.985	**0.989**	**0.928**	**0.968**	**0.993**	**0.993**
GFT (Oct 2014)	0.875	**0.989**	0.968	0.833	0.926	0.969	0.986
Ref. 16	0.971	0.967	0.983	0.927	0.956	0.985	0.984
GFT+AR(3)	0.967	0.986	0.985	0.879	0.929	0.945	0.957
AR(3)	0.964	0.968	0.971	0.877	0.903	0.927	0.945
Naive	0.961	0.951	0.954	0.887	0.924	0.923	0.937
Correlation of increment							
ARGO	**0.758**	0.806	**0.810**	0.286	**0.527**	**0.938**	0.912
GFT (Oct 2014)	0.706	**0.863**	0.702	0.484	0.502	0.847	**0.918**
Ref. 16	0.690	0.776	0.693	**0.510**	0.367	0.915	0.889
GFT+AR(3)	0.512	0.708	0.708	0.165	0.141	0.534	0.587
AR(3)	0.385	0.585	0.569	0.077	0.011	0.404	0.493
Naive	0.436	0.602	0.570	0.095	0.134	0.406	0.514

GFT+AR(3) stands for the model $p_t = \mu + \alpha_1 p_{t-1} + \alpha_2 p_{t-2} + \alpha_3 p_{t-3} + \beta \mathrm{GFT}(t)$, where the GFT estimate is treated as an exogenous variable. Boldface highlights the best performance for each metric in each study period. RMSE, MAE, and MAPE are relative to the error of naive method; that is, the number reported is the ratio of error of a given method to that of the naive method. The absolute error of the naive method is reported in parentheses. All comparisons are based on the original scale of ILI activity level.

$$y_t = \mu_y + \sum_{j=1}^{N} \alpha_j y_{t-j} + \sum_{i=1}^{K} \beta_i X_{i,t} + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

Combines:

- lots of lags (statistical issue or fundamental model?)
- lots of search terms
- regularization (different way of dealing with over-fitting)

Advances in nowcasting influenza-like illness rates using search query logs

Vasileios Lampos[1,2], Andrew C. Miller[2,3], Steve Crossan[2] & Christian Stefansen[2]

Lampos et al 2015 "Advances in nowcasting influenza-like illness rates using search query logs" *Scientific Reports*, `http://dx.doi.org/10.1038/srep12760`

- Eysenbach (2006)
- Polgreen et al (2008)
- Ginsberg et al (2009)
- Goel et al (2010)
- Cook et al (2011)
- Lazer et al (2014)
- Yang et al (2015), Lampos et al (2015)

Can you think of an area in your field that has made this much progress?