CS 8803 Data Analytics for Well-being: Data Modeling II

Munmun De Choudhury

munmund@gatech.edu

Week 4 | February 3, 2016
Replacement Class

- At CSCW 2016 on Feb 29 and Mar 2.
- One replacement class: GVU Brown Bag talk tomorrow – distinguished speaker of the semester.
  - TSRB Ballroom 11:30 – lunch available; 12 noon-1pm – talk; 1-1:30pm – student Q&A
  - Talk title: *Online Social Support: Advances in Measuring Support and Understanding Its Effects*

![Robert E. Kraut](image)

*Robert E. Kraut*

*Home Page*

Herbert A. Simon Professor of Human-Computer Interaction
Human-Computer Interaction Institute, School of Computer Science
Tepper School of Business
Center for the Future of Work, Heinz School
Carnegie Mellon University
Pittsburgh, PA 15213

*Home*

*Research*
Detecting influenza epidemics using search engine query data
Main Idea

- The academic paper on Google Flu Trends: it relied on trends in search terms, such as headache and chills, to estimate the number of flu cases.
- The search terms were correlated with flu outbreak data collected by the Centers for Disease Control and Prevention (CDC).
Method

- By aggregating historical logs of online web search queries submitted between 2003 and 2008, the authors computed time series of weekly counts for 50 million of the most common search queries in the US.
- The authors built a linear model using the log-odds of an influenza-like illness (ILI) physician visit (CDC data from nine regions) and the log-odds of an ILI-related search query.

\[ \text{logit}(P) = \beta_0 + \beta_1 \times \text{logit}(Q) + \varepsilon \]
42 points obtained a mean correlation of 0.97 (min=0.92, max=0.96, n=9 regions) (Figure 2). CDC does not make state-level data publicly available, but we can be processed quickly, the resulting ILI estimates were excluded from all prior steps. Estimates generated for these regions, we were able to consistently estimate the current ILI percentage 1-2 weeks ahead of the publication of reports. Other search queries in the top 100, not included like illnesses. Other search queries in the top 100, not included

Throughout the 2007-2008 influenza season, we used Using this ILI-related query fraction as the explanatory filter. Billions of queries occurred infrequently and were temporally

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

<table>
<thead>
<tr>
<th>Search Query Topic</th>
<th>Top 45 Queries</th>
<th>Next 55 Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Weighted</td>
<td>N</td>
</tr>
<tr>
<td>Influenza Complication</td>
<td>11 18.15</td>
<td>5   3.40</td>
</tr>
<tr>
<td>Cold/Flu Remedy</td>
<td>8   5.05</td>
<td>6   5.03</td>
</tr>
<tr>
<td>General Influenza Symptoms</td>
<td>5   2.60</td>
<td>1   0.07</td>
</tr>
<tr>
<td>Term for Influenza</td>
<td>4   3.74</td>
<td>6   0.30</td>
</tr>
<tr>
<td>Specific Influenza Symptom</td>
<td>4   2.54</td>
<td>6   3.74</td>
</tr>
<tr>
<td>Symptoms of an Influenza Complication</td>
<td>4   2.21</td>
<td>2   0.92</td>
</tr>
<tr>
<td>Antibiotic Medication</td>
<td>3   6.23</td>
<td>3   3.17</td>
</tr>
<tr>
<td>General Influenza Remedies</td>
<td>2   0.18</td>
<td>1   0.32</td>
</tr>
<tr>
<td>Symptoms of a Related Disease</td>
<td>2   1.66</td>
<td>2   0.77</td>
</tr>
<tr>
<td>Antiviral Medication</td>
<td>1   0.39</td>
<td>1   0.74</td>
</tr>
<tr>
<td>Related Disease</td>
<td>1   6.66</td>
<td>3   3.77</td>
</tr>
<tr>
<td>Unrelated to Influenza</td>
<td>0   0.00</td>
<td>19  28.37</td>
</tr>
<tr>
<td></td>
<td><strong>45 49.40</strong></td>
<td><strong>55 50.60</strong></td>
</tr>
</tbody>
</table>

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, while a correlation of 0.96 was obtained over 42 validation points. 95% prediction intervals are indicated.
What happened later...

• Initially Google’s tracker appeared to be pretty good, matching CDC data’s late-breaking data somewhat closely.

• Two notable stumbles led to its ultimate downfall:
  • an underestimate of the 2009 H1N1 swine flu outbreak
  • an alarming overestimate (almost double real numbers) of the 2012-2013 flu season’s cases

• But what happened really?
One of the fatal problems with the model was that it didn’t account for shifts in people’s search behavior.

For instance, in the 2012-2013 flu season many people were searching for news about the flu season, rather than plugging in flu symptoms for diagnostics.

Over time, people also use different terms to search for the same things.
Discovering Health Topics in Social Media Using Topic Models
Main Idea

- The article proposes an unsupervised machine learning methodology to discover ailment topics in Twitter content.
- The authors specifically present a statistical topic model, the Ailment Topic Aspect Model (ATAM) for the purpose — it combines a background model, a topic model and an ailment model with word (WebMD) and topic (ailment) priors.
- Results demonstrate that it is possible to automatically discover topics that attain statistically significant correlations with ground truth data, despite using minimal human supervision and no historical data to train the model.
Each word is conditionally independent given the parameters. LDA is a Bayesian model in which there are also distributions (priors) over the parameters $h$ and $w$, given by Dirichlet distributions with hyperparameters $a$ and $b$.

In our experiments, we use a variant of LDA that includes an additional ''background'' word distribution to model common, non-topical words, which can produce less noisy topics. [36–37] This model assumes that each word is generated under the standard LDA model with probability $l$, while with probability $1-l$ the word comes from the background distribution. This concept is also in ATAM, described below.

Figure 3. Influenza over time. The weekly rate of influenza as estimated by the volume of tweets assigned to the influenza-like illness topics and keywords alongside the rates given by the CDC ILINet (solid black line). The better of the two LDA topics is shown. All rates are standardized (z-scores) so that they are comparable on the y-axis.

doi:10.1371/journal.pone.0103408.g003

Table 1. Pearson correlations between various Twitter models and keywords and CDC influenza-like illness (ILI) surveillance data for three time periods.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ATAM</td>
<td>.613</td>
<td>.643</td>
<td>.689</td>
</tr>
<tr>
<td>LDA (1)</td>
<td>.670</td>
<td>.198</td>
<td>.455</td>
</tr>
<tr>
<td>LDA (2)</td>
<td>2.021</td>
<td>.698</td>
<td>.637</td>
</tr>
<tr>
<td>''flu''</td>
<td>.259</td>
<td>.652</td>
<td>.717</td>
</tr>
<tr>
<td>''influenza''</td>
<td>.509</td>
<td>.767</td>
<td>.782</td>
</tr>
</tbody>
</table>

The two LDA rows correspond to two different LDA topics.

doi:10.1371/journal.pone.0103408.t001
Ailment Topic Aspect Model (ATAM).

Preliminary LDA experiments discovered health-related topics around ailments but many other topics as well. For example, some topic clusters would correspond to symptom terms that could be associated with many illnesses.

Consider the example sentence, "damn flu, home with a fever watching TV." It contains two words relevant to the ailment of flu (''flu,'' ''fever''), one of which is a symptom. It also contains words that are not about the ailment but are topically related (''home,'' ''watching,'' ''TV''), which might be described by a ''stay at home'' topic. Finally, it contains common words that would not be described with a particular topic or ailment (''damn,'' ''with,'' ''a'').

We developed a model that explicitly labels each tweet with an ailment category and distinguishes ailment words from other topics and non-topical words. Our model includes a standard LDA.
Twitter is used by millions, but could it also have bias in such measurements?
The first paper focuses on flu and the second on a variety of chronic or seasonal illnesses. Are there conditions that will be hard to capture through search logs or social media?
Could alternative social media platforms help us understand some of these other less socially sharable illnesses?
The topic modeling approach improves on simple techniques like those using dictionaries. What are its limitations?
Could there be latent factors that drive sharing of ailment related topics on social media?
People use social media for all kinds of reasons and purposes. Would that affect the moods they express?
Would “self-presentation”, “social comparison” or identity impact the kinds of illnesses shared?
The ATAM model paper does not show that the illness trends precede the CDC or the Gallup measures. How can one leverage these social media based measurements?