Use of Social Media to Monitor and Predict Outbreaks and Public Opinion on Health Topics

Alessio Signorini

Department of Computer Science
University of Iowa

December 3rd, 2014
“Measurement is the first step that leads to control and eventually to improvement.”

- James Harrington
Data Analytics

- Nascar / Formula One
- Sports
- Insurances
- Sales / Marketing
- Online Advertising
- Logistics
in Public Health we have

Disease Surveillance
Surveillance Systems

- Vital Statistics & Registries (e.g., births, deaths, defects)
- Population Surveys (e.g., substance abuse)
- Disease Reporting (e.g., salmonellosis, measles)
- Sentinel Surveillance (e.g., Influenza-Like Illnesses)
- Adverse Events Surveillance (e.g., issues with drugs)
- Laboratory Data
surveillance data should be a byproduct of any healthcare operation
Syndromic Surveillance

- Focuses on Early Detection
- Based on disease signs or symptoms, not diagnosis
- Novel sources: Emergency Room data, Drugs sales
- Uses well known Data Mining techniques
- Reduced delay in results
aggregate and analyze
Social Media Data
to monitor and predict health trends
Online:
~27h/mon

Mobile:
~34h/mon

- Social: 22%
- Content: 19%
- Email/IM: 21%
- Video: 20%
- Shopping: 13%
- Search: 5%

Platforms:
- Facebook: ~5B/day
- Twitter: ~500M/day
- Foursquare: ~7M/day
- WordPress: ~10M/day
Google Searches

Monitor
Public Opinion

Positive Tweets

Comprehensive Exam
Alessio Signorini
University of Iowa, May 2010
The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic

Alessio Signorini, Alberto Segre, Philip Polgreen

Using Twitter to Estimate H1N1 Activity
Alessio Signorini, Alberto Segre, Philip Polgreen
ISDS 2010 – 9th Annual Conference of International Society for Disease Surveillance
Inferring Travel from Social Media
Alessio Signorini, Alberto Segre, Philip Polgreen
ISDS 2011 – 10th Annual Conference of International Society for Disease Surveillance

Monitor Travels

National

Local
can we use “Social Travel Models” to improve local flu trends prediction?
City-Level Flu Trends

- CDC’s MMWR - Flu & Pneumonia Deaths for 122 cities
- Smoothed each week with values of prev/next 2 weeks
Social Travel Data

- 240 Million geolocated tweets posted by 4 Million users
- Mapped over MMWR cities, discarded overlapping ones
- Used Spark cluster of 8 machines to do geo-mapping

![Graph showing volume of trips among MMWR cities in 2012](image)
Social Travel Model

- **Final dataset**: 78 cities, 124M tweets, 2.2M users
- Assumed “home” the most common location
- A “trip” was a post at home followed by one elsewhere
- Used population to scale volume of trips between cities

<table>
<thead>
<tr>
<th>City</th>
<th>Population</th>
<th>Home for</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York City, NY</td>
<td>8406000</td>
<td>297486</td>
<td>3.54%</td>
</tr>
<tr>
<td>Los Angeles, CA</td>
<td>3857799</td>
<td>206378</td>
<td>5.35%</td>
</tr>
<tr>
<td>Chicago, IL</td>
<td>2714856</td>
<td>94952</td>
<td>3.50%</td>
</tr>
<tr>
<td>Houston, TX</td>
<td>2160821</td>
<td>71780</td>
<td>3.32%</td>
</tr>
<tr>
<td>Philadelphia, PA</td>
<td>1547607</td>
<td>77626</td>
<td>5.02%</td>
</tr>
</tbody>
</table>
Correlation b/w Cities

San Jose, CA
Atlanta, GA
Philadelphia, PA
Predicting Flu Trends

- Flu Trends of 78 cities generated from MMWR data
- Used 2011 for training and 2012 for testing
- Support Vector Regression with polynomial kernel
- **Target**: value of local flu trend for that week
- **Features**: value of top 20 correlated cities 2 weeks before
Measures Compared

- **Distance**: closest 20 cities
- **Similarity**: most similar 20 cities on 2011 flu trends
- **Flow**: top 20 cities by number of visitors

<table>
<thead>
<tr>
<th></th>
<th>Distance</th>
<th>Similarity</th>
<th>Flow</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0.23086</td>
<td>0.26971</td>
<td>0.28080</td>
<td>0.28543</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.00009</td>
<td>0.00004</td>
<td>0.00027</td>
<td>0.00001</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.7780</td>
<td>0.73145</td>
<td>0.83247</td>
<td>0.79387</td>
</tr>
</tbody>
</table>

Table 3.4: Square Correlation Coefficients for each approach
Prediction Results

<table>
<thead>
<tr>
<th>City</th>
<th>Distance</th>
<th>Similarity</th>
<th>Flow</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Washington, DC</td>
<td>0.043178</td>
<td>0.066362</td>
<td>0.013697</td>
<td>0.001167</td>
</tr>
<tr>
<td>Philadelphia, PA</td>
<td>0.000334</td>
<td>0.057870</td>
<td>0.007827</td>
<td>0.002519</td>
</tr>
<tr>
<td>Dallas, TX</td>
<td>0.000088</td>
<td>0.024712</td>
<td>0.005816</td>
<td>0.001576</td>
</tr>
<tr>
<td>Waterbury, CT</td>
<td>0.061919</td>
<td>0.004790</td>
<td>0.002763</td>
<td>0.00006</td>
</tr>
<tr>
<td>Fort Wayne, IN</td>
<td>0.005689</td>
<td>0.000038</td>
<td>0.000272</td>
<td>0.000518</td>
</tr>
</tbody>
</table>

Table 3.6: Square Correlation Coefficients for most difficult cities

<table>
<thead>
<tr>
<th>City</th>
<th>Distance</th>
<th>Similarity</th>
<th>Flow</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Las Vegas, NV</td>
<td>0.757804</td>
<td>0.562152</td>
<td>0.83247</td>
<td>0.735552</td>
</tr>
<tr>
<td>San Jose, CA</td>
<td>0.602464</td>
<td>0.689917</td>
<td>0.801097</td>
<td>0.793866</td>
</tr>
<tr>
<td>Albuquerque, NM</td>
<td>0.556896</td>
<td>0.626601</td>
<td>0.711688</td>
<td>0.494305</td>
</tr>
<tr>
<td>Tucson, AZ</td>
<td>0.740066</td>
<td>0.5472</td>
<td>0.694016</td>
<td>0.661322</td>
</tr>
<tr>
<td>San Antonio, TX</td>
<td>0.467228</td>
<td>0.636067</td>
<td>0.686346</td>
<td>0.612461</td>
</tr>
</tbody>
</table>

Table 3.5: Square Correlation Coefficients for most predictable cities
Failure Hypothesis

- Port-of-entry influenced by international travels
- Noisy data: Watebury, CT had only 43 deaths in 2011
- Few data: Fort Wayne has 1/50th of Las Vegas’ users

[Graph showing flu deaths in Washington, DC from 2012]
Conclusions

- Social Media can be an important source for surveillance
- Can predict American Idol’s winner ;)
- Allows to monitor public sentiment about health topics
- Can effectively be used to monitor ILI% in real time
- Geolocated posts can be used to create travel models
- Social Travel Data provides additional predictive power for flu trends
Checkins Distributions

![Graph showing checkins distributions.]

- **0 < 1 mile**: 50%
- **1 < 10 miles**: 85%
- **10 < 100 miles**: 97%
- **100 < 1000 miles**: 99%
- **1000 < 10000 miles**: 100%

- **% Trips**
- **% Cumulative**

![Graph showing time distribution.]

- **10s**: 0%
- **30s**: 1%
- **1m**: 4%
- **2m**: 8%
- **5m**: 15%
- **10m**: 21%
- **15m**: 24%
- **30m**: 31%
- **1h**: 38%
- **2h**: 46%
- **6h**: 59%
- **12h**: 69%
- **1d**: 81%
- **2d**: 89%
- **1w**: 97%

- **% Trips**
- **% Cumulative**
Denver, CO

Distance

Similarity

Flow
Smoothing Methods

5 weeks ahead

1 week around

2 weeks around