Language Modeling for Places from Social Media

Vanessa Murdock
Microsoft
51° 30' 50.0868", 0° 7' 42.8514"

Photo Courtesy NASA
But where are we?
What are we doing?

Photo by threat to democracy on Flickr http://www.flickr.com/photos/16725630@N00/3974387517
Why?

• We can leverage this information to create applications and services that are relevant, engaging and delightful
• User context is required to provide minimally relevant results
• Mobile devices are increasingly the primary portal through which users seek information
  • Mobile devices are personal, less likely to be shared
  • Increasingly “intelligent” raising users’ expectations of the device’s understanding of context
• Social media can provide information about where a user is, and what they are thinking, feeling and doing in that place
  • Often more direct than search query logs or browser logs
Local Search Dependent on Local Data

Yelp is the best way to find great local businesses. People use Yelp to search for everything from the city's tastiest burger to the most renowned cardiologist. What will you uncover in your neighborhood?

Create Your Free Account

Best of Yelp: London

Review of the Day

Nicole R.

I really like the Melting Pot. It's fantastic. She recommended Pittsburgh's so we went there for a Saturday night. It was packed, but we could have a seat by the bar and the service was excellent. The appetizers were delicious and the prime rib was cooked to perfection.

Foursquare helps you find the perfect places in London to go with friends:
One Strategy is to Own and Curate

- **Pros:**
  - Content can be controlled
  - Clean data
- **Cons:**
  - Expensive to curate and maintain
  - Limited scope
  - Single data source
…Or Let the Users Do the Work

- Pros:
  - Data can be rich in attributes
  - Deep coverage
    - Naming variants
    - Off-the-beaten track
  - Cheap to acquire (once the App is adopted)

- Cons:
  - Data is messy
  - Difficult to get users to adopt the App
  - Single source of data (your App)
    - Potentially biased by a specific set of users
...Or Extract it From Social Media

• Pros:
  – Data can be rich in attributes
  – Deep coverage
    • Naming variants
    • Off-the-beaten track
  – Little investment in data curation
  – Multiple complementary sources

• Cons:
  – Data is messy
  – Licenses may be onerous
  – Conflation is difficult
  – Information must be extracted
About Social Media Data

• Generated in unprecedented quantities
• Biased toward the “Technorati”
  – This is changing as smart phones become cheaper and more widely adopted
• Often generated by people on vacation
  – Biased toward known tourist destinations
  – Tagged by people not intimately familiar with the location
• Assumes the user’s context is known
  – Location implicit
  – Content meaningful only to the user’s social circle
• Place names suggested to the user
  – Reduces the richness of the vocabulary
35m of 130m Geotagged Images in Flickr

David Crandall et al., Cornell University
Geographic Distribution of 4SQ Checkins
Modeling Locations
Lay a grid over the globe
Associate social media items with each “cell” according to its geotags.
Estimate a term distribution from each (cell) "document"
Do information retrieval
Jolie gare des années 1920. Influence angkorienne et ressemblance frappante avec le palais du roi Narai, à Lopburi... Nice, small Paris railway station from the 1920’s. Through the Angkorian influence, strong similarities with the King Narai’s Palace, in Lopburi...
Jolie gare des années 1920. Influence angkorienne et ressemblance frappante avec le palais du roi Narai, à Lopburi… Nice, small Paris railway station from the 1920’s. Through the Angkorian influence, strong similarities with the King Narai’s Palace, in Lopburi…

Tags:
gare
pont
cardinet
paris
RER
1920
1925
Angkor
ankorien
Angkorian
Lopburi
Narai
pluie
rain
reflet
reflection
zebra
Compute the probability a text was generated by a cell
Rank locations by the probability score
Term Statistics vs. User Statistics

\[ P(t \mid L) = \frac{c(t;L)}{|L|} \]

Number of times the term is used in a location divided by the number of terms in the location

\[ P(t \mid U,L) = \frac{c(u,t,L)}{\sum_{t_i \in L} c(u,t_i,L)} \]

Number of users tagging with a term in a location divided by the user frequency for any term in that location
Results: Predicting the location of a Flickr photo from its tags

<table>
<thead>
<tr>
<th></th>
<th>Acc</th>
<th>Acc@1</th>
<th>Acc@2</th>
<th>Acc@3</th>
</tr>
</thead>
<tbody>
<tr>
<td>100KM</td>
<td>0.587</td>
<td>0.670</td>
<td>0.698</td>
<td>0.718</td>
</tr>
<tr>
<td>10KM</td>
<td>0.394</td>
<td>0.516</td>
<td>0.546</td>
<td>0.563</td>
</tr>
<tr>
<td>1KM</td>
<td>0.172</td>
<td>0.311</td>
<td>0.364</td>
<td>0.397</td>
</tr>
</tbody>
</table>

Specific to social media:
Term distributions best estimated by user frequency, rather than term frequency
Test/train split must be done on users, not on photos
There goes our Montreal trip and NJ concert. RT @u2gigs: U2's North American dates will be rescheduled in 2011 http://u2.gs/5K [40.735048,-73.98637]

I'm at K & M Camera (385 Broadway, Walker and White, New York). http://4sq.com/8Bc01h [40.718381,-74.003091]

lol agree but I got next RT @ubrokebird: @NoelElMagnifico @ULuvRidinHeidi @sextoydiva im thinkiin if its mud (cont) http://tl.gd/1fslan  [40.752321,-73.985531]

Mayor @mikebloomberg's on stage at #tcdisrupt talking smack about the bay area for startups. Where are you @gavinnewsom? [40.72531753,-74.01140505]
Modeling Locations with Tweets

- Tweets are closer to natural language
  - Reflect everyday lives
  - Flickr often reflects vacation lives

- Same basic approach but with WOEIDs instead of grid cells
  - Closer to the “real world”
  - Hierarchical, fewer arbitrary boundaries
  - Comparison to Placemaker
Placemaker

• Toolbox for extracting and disambiguating locations in text
  • Rule-based system relying on gazetteers
• Includes functionality for defining the geographic scope of a document
• Originally a free service developed at Yahoo, now part of their paid services (BOSS)
• Populated with licensed data from multiple sources
  – Includes language variants, colloquial names
Results: Tweet Location Prediction

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Country</th>
<th>State</th>
<th>City</th>
<th>Zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placemaker</td>
<td>0.528</td>
<td>0.407</td>
<td>0.269</td>
<td>0.017</td>
</tr>
<tr>
<td>Language Model</td>
<td>0.532</td>
<td>0.316</td>
<td>0.298</td>
<td>0.139</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accuracy@2</th>
<th>State</th>
<th>City</th>
<th>Zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placemaker</td>
<td>0.462</td>
<td>0.342</td>
<td>0.029</td>
</tr>
<tr>
<td>Language Model</td>
<td>0.458</td>
<td>0.326</td>
<td>0.188</td>
</tr>
</tbody>
</table>
## Results: User Location Prediction

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Country</th>
<th>State</th>
<th>City</th>
<th>Zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placemaker</td>
<td>0.577</td>
<td>0.471</td>
<td>0.314</td>
<td>0.023</td>
</tr>
<tr>
<td>Language Model</td>
<td>0.759</td>
<td>0.449</td>
<td>0.319</td>
<td>0.135</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accuracy@2</th>
<th>State</th>
<th>City</th>
<th>Zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placemaker</td>
<td>0.518</td>
<td>0.379</td>
<td>0.038</td>
</tr>
<tr>
<td>Language Model</td>
<td>0.589</td>
<td>0.362</td>
<td>0.213</td>
</tr>
</tbody>
</table>
Case Study: Disambiguating POIs

• News photos to illustrate a story
  – Caption, title and location provided with the data
• Captions assigned by journalists or editors
• Locations extracted from the caption automatically prior to publication
• 74 examples of points of interest
• Ground truth created by looking up each POI on a map and identifying its coordinates
  – Visually verified in satellite view
  – Those that could not be verified were discarded
Three Experiments

• Placemaker was the state of the art, and the baseline for predicting POI locations
• Use the highest precision location model to predict where the POI is
• Use the two techniques together in a cascade
  – If the POI is found by Placemaker, use it
  – Otherwise predict the location with the model
• Evaluation: Median distance from true location
## Results

<table>
<thead>
<tr>
<th></th>
<th>All Data 74 examples</th>
<th>PM POIs 23 examples</th>
<th>Other POIs 51 examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median</td>
<td>Mean</td>
<td>Median</td>
</tr>
<tr>
<td>Placemaker</td>
<td>7.0</td>
<td>115.8</td>
<td>0.233</td>
</tr>
<tr>
<td>LM</td>
<td>0.469</td>
<td>14.8</td>
<td>0.522</td>
</tr>
<tr>
<td>Cascade</td>
<td>0.322</td>
<td>11.0</td>
<td>0.233</td>
</tr>
</tbody>
</table>

- **PM POIs:** POIs as identified by Placemaker
- **74 POIs corresponds to 6000 images**
- **A few bad guesses biases the mean**
“Spanish F1 Grand Prix”
Algorithm: Truncate and Aggregate

Initialize k;
N = 4;
DO:
    Round Latitude/Longitude to N places (a “cell”);
    Aggregate social media data with same “cell”;
    Count the vocabulary V;
    If (|V| > k) store the “cell”;
    Else N = N – 1;
WHILE (N >= 0)
Dynamically Location Models

• Physical size of the cell dependent on the underlying data
• Places represented with the smallest granularity supported by the data
  • Allows prediction of points of interest very accurately
  • Sparsely represented places will not be accurate, regardless of the granularity, so backing off to a larger area with more data is the only sensible option
• Low variance in the term distributions
  • Prior on the location encoded in the model
• Lends itself to privacy preservation
  • Discard terms used by fewer than k users
  • Discard places represented by fewer than k people
### Evaluation: Predicting location of Flickr Tag Sets

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>1 decimal</th>
<th>2 decimals</th>
<th>3 decimals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Median</strong></td>
<td>14.29</td>
<td>14.71</td>
<td>17.24</td>
<td></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>1546</td>
<td>1665</td>
<td>1695</td>
<td></td>
</tr>
<tr>
<td><strong>Vocab</strong></td>
<td>50</td>
<td>100</td>
<td>250</td>
<td></td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>4.5</td>
<td>4.48</td>
<td>5.89</td>
<td></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>1276</td>
<td>1157</td>
<td>972</td>
<td></td>
</tr>
<tr>
<td><strong>Users</strong></td>
<td>5</td>
<td>20</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>4.26</td>
<td>5.37</td>
<td>8.30</td>
<td></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>1227</td>
<td>1001</td>
<td>1066</td>
<td></td>
</tr>
</tbody>
</table>
Papers and Links

• Eric Fisher: Locals and Tourists  
  https://www.flickr.com/photos/walkingsf/sets/72157624209158632/
• David Crandall: Mapping the World’s Photos  
  http://www.cs.cornell.edu/~crandall/photomap/
• Tom Taylor: Neighborhood Boundaries  http://boundaries.tomtaylor.co.uk/
• Modeling Locations with Social Media:  
• Modeling Locations with Tweets:  http://doras.dcu.ie/16754/1/nohare_paper.pdf
• Mining the Web for Points of Interest:  
  http://cs.uef.fi/pages/franti/lami/papers/Mining%20the%20Web%20for%20Points%20of%20Interest.pdf
• Dynamic Location Models:  http://dl.acm.org/citation.cfm?id=2609552
Thank you!