Harvesting Geospatial Knowledge from Social Metadata

Suradej Intagorn
Anon Plangprasopchok
Kristina Lerman

USC Information Sciences Institute
Motivation

Tags:
- station
- fire
- los
- angeles
- california
- satellite
- image
- digitalglobe
- Los Angeles
- USA

Credit: DigitalGlobe
Goals

Problem
- How to extract *geospatial knowledge* from user contributed annotations (metadata)?

Advantages
- Cheaper and Up-to-date than knowledge produced by a group of experts
- Supplement to exist knowledge created by experts
Overview of Approach

Combine geospatial annotations from many users to learn more complete representation of places & their relations

• tags,
• geo-tags
• set names (to be explained)

Challenges:
1. Place names are often ambiguous
2. Noisy data points
3. Sparseness
Types of Social Meta Data

The old parola of Opon probably dates back to the 50s. It is no longer in operation but remains a tourist attraction.

Tags:
- Opon
- Mactan
- Lapu-Lapu
- Church
- Cebu
- Philippines
- archway
- jetty
- port
- dock
- red
- white
- colors
- parola
- lighthouse
- faro
- Cebu-Sugbo

Additional Information:
- All rights reserved
- Anyone can see this photo
- Taken in Lapu-Lapu City
- Taken on August 4, 2005
- 3 people call this photo a fail
- Viewed 767 times

This photo also belongs to:
- aPeeling Faint (Pool)
- PinoyS Galore (Pool)
- Philippines (Pool)
- LIGHHOUSE LOVERS (Pool)
- SoutheastAsia Images (Pool)
- Philippines Images (Pool)
- Historic Preservation (Pool)
- Catchy Colors (Pool)
- "Only in the Philippines!" (Pool)
- visit the world - the travel guide (Pool)
- Intense and Vibrant colors (Pool)
- Travel Photography (Pool)
Types of Social Meta Data
Types of Social meta data
### Summary of Social Metadata

- Each photo is just a point in spatial domain.
- Each set is a collection of points representing a place.
- They can be aggregated to get better representation of the place.

<table>
<thead>
<tr>
<th>Photo</th>
<th>Photo ID</th>
<th>User ID</th>
<th>Set ID</th>
<th>Set Name</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
</table>
Approach

- Recognizing Places
- Disambiguating Places
- Noise Filtering
- Representing Places
- Geographic subsumption

GeoNames
Approach

Recognizing Places

Disambiguating Places

Noise Filtering

Representing Places

Geographic subsumption

GeoNames
Recognizing Place-related Set Names

Assumption

- Set names refer to places
- But, need to filter out non-place set name

Current approach

- Use reference set from GeoNames.org
- Simple name matching with set name
Approach

Recognizing Places

Disambiguating Places

Noise Filtering

Representing Places

Geographic subsumption

GeoNames
Challenges: Ambiguity

Place names are often ambiguous

Victoria ?

Victoria, BC, Canada

Victoria, Australia
Disambiguating Places

Problem: Place may be also non-contiguous
Two “USA” places: (1) United States; (2) Usa, Japan
Disambiguating Places

Assumption: ambiguous places can be disambiguated by their locations

How?
1. Construct graph of geo points
2. Then, finding connected components
Learning Non-Contiguous Places
Disambiguating Places

Construct Graph $G_1 = (V, E)$,
- Vertices, $V$, are points corresponding to geo-tagged photos
- Edges, $E$, are the edges between vertices

When is the edge between two photos created?

$$\text{dist}(v_i, v_j) < \tau \text{ or } s_{v_i} = s_{v_j}.$$
Multi Cluster Representation

USA

Semantically disambiguated places

USA (United States)

- Hawaii cluster
- East cluster
- West cluster

Usa, Japan

- Usa, Japan Cluster

Cluster representation level
Approach

- Recognizing Places
- Disambiguating Places
- Noise Filtering
- Representing Places
- Geographic subsumption
Heuristic: Cluster that contains photos from < k users is noise
Approach

- Recognizing Places
- Disambiguating Places
- Noise Filtering
- Representing Places
- Geographic subsumption
Representing Places

Multiple Convex Hulls
Learning relations between places
Geographic Subsumption

Based on probabilistic subsumption
(Sanderson and Croft, 1999; Schmitz, 2006)

place \( A \) subsumes place \( B \) if
\[
p(A|B) \geq t \text{ and } p(B|A) < t
\]

, where \( t \) is a predefined threshold ranging from 0 to 1

\[
p(A|B) = \frac{\text{Area}(A \cap B)}{\text{Area}(B)}
\]
\[
p(B|A) = \frac{\text{Area}(A \cap B)}{\text{Area}(A)}
\]
Probabilistic Subsumption on Tags

Probabilistic subsumption on tags (Schmitz, 2006)

place A subsumes place B if
\[ p(A|B) \geq t \text{ and } p(B|A) < t \]

\[ p(A|B) = \frac{\text{Frequency}(A,B)}{\text{Frequency}(B)} \]
\[ p(B|A) = \frac{\text{Frequency}(A,B)}{\text{Frequency}(A)} \]

*Frequency*(A,B): how many times tags A and B co-appear in the data set;
*Frequency*(A): how many times tag A appears in the data set.
Result: # of Induced Relations

- Geo Subsumption: All Continents
- Geo Subsumption: North America
- Tag Subsumption: North America
Evaluate the quality of learned relations against the reference set from GeoNames.org
Result: Quality of the Learned Relations

\[ F\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]
### Examples of Novel Relations

<table>
<thead>
<tr>
<th>Child</th>
<th>Parent</th>
<th>Child</th>
<th>Parent</th>
</tr>
</thead>
<tbody>
<tr>
<td>anaheim</td>
<td>la</td>
<td>disneyland resort</td>
<td>disneyland</td>
</tr>
<tr>
<td>ballard</td>
<td>puget sound</td>
<td>disneyland</td>
<td>la</td>
</tr>
<tr>
<td>brandywine park</td>
<td>wilmington</td>
<td>golden gate bridge</td>
<td>san francisco bay</td>
</tr>
<tr>
<td>bronx</td>
<td>new york city</td>
<td>pearl harbor</td>
<td>oahu</td>
</tr>
<tr>
<td>bronx zoo</td>
<td>new york city</td>
<td>times square</td>
<td>new york city</td>
</tr>
<tr>
<td>coney island</td>
<td>new york city</td>
<td>university of washington</td>
<td>puget sound</td>
</tr>
</tbody>
</table>
Related Work

- Inducing conceptual hierarchies from tags’ frequencies [Schmitz, 2006; Brooks & Montanez, 2006, Mika, 2007]
- Inducing conceptual hierarchies from collection – set relations [Plangprasopchok & Lerman, 2009]
- Identifying “place tags” from geopoints [Rattenbury & Naaman, 2009]
- Web-A-Where [Toni: please shortly mention it]
- Li et al. [Toni: please shortly mention it]
- Spirit tagger?
Our approach can:

- learn more high quality relations than the baseline
- learn novel relations that do not exist in expert-created knowledge bases
- leverage user generated knowledge to aid decision makers
Thank you