Latent Geographic Feature Extraction from Social Media

Christian Sengstock*
Michael Gertz

Database Systems Research Group
Heidelberg University, Germany

November 8, 2012
Social Media is a huge and increasing source of unstructured and uncertain geographic information.

Effort to make data usable:

- (Structured) Information Extraction
  Place/event extraction from Flickr [Rattenburry SIGIR’07]
  Event trajectory extraction from Twitter [Sakaki WWW’10]

- Spatial Analysis
  Spatio-temporal forecasting using Flickr [Jin MM’12]
  Study ecological phenomena [Zhang WWW’12]
**General Motivation:** Extract spatial variables from unstructured and noisy geographic information sources

**This work:** Framework for unsupervised extraction of informative spatial variables (dimensions of geographic semantics) from Social Media
Outline

1. Definitions and Problem Statement
2. Data Characteristics and Normalization
3. Latent Geographic Feature Extraction
4. Experiments
5. Conclusions
Outline

1. Definitions and Problem Statement
   - Geographic Feature
   - Signal Estimation
   - Problem Statement

2. Data Characteristics and Normalization
   - Distribution Characteristics
   - Normalization
   - Geographic Feature Types

3. Latent Geographic Feature Extraction
   - Dimensionality Reduction
   - Framework

4. Experiments
   - Technique Comparison
   -Normalization Influence
   - Exploration Task

5. Conclusions
Terminology

- **Geographic Feature $f$**
  - A dimension representing some semantics of a location (e.g., temperature, population, number of restaurants)
  - Sampled (measured) at any location $l$ in geographic space $W$ ($\rightarrow$ spatial variable)

- **Geographic Feature Sensor $\phi$ and Signal $\phi(l)$ of $f$:**

\[
\phi : W \rightarrow \mathbb{R}_+
\]

\[
l \mapsto \phi(l)
\]
Terminology

- Set of geographic features $f_1, \ldots, f_p$ defines a **Multivariate Geographic Feature Sensor**:
  \[
  \Phi := (\phi_1, \ldots, \phi_p)^T
  \]

- Spatial sampling scheme (measurements) $L = (l_1, \ldots, l_n)$ defines a **Location Sampling Matrix**:
  \[
  X^{n \times p} = (\Phi(l_1), \ldots, \Phi(l_n))^T = \\
  \begin{pmatrix}
  \phi_1(l_1) & \cdots & \phi_p(l_1) \\
  \vdots & \ddots & \vdots \\
  \phi_1(l_n) & \cdots & \phi_p(l_n)
  \end{pmatrix}
  \]
A Social Media Collection $D$ consists of documents:

$$d_i = (X, u, l, t)$$

- $X$: Bag of document features (terms, tags, image features, ...)
- $u$: User
- $l$: Location
- $t$: Timestamp

**Assumption:** Features with geographic meaning aggregate in subsets of geographic space $\rightarrow$ high signal
Signal Estimation

- Every **document feature** \( f_1, \ldots, f_p \) is a possibly meaningful/meaningless **geographic feature**

- Intuition of geographic feature signal \( \phi_i(l) \): 
  
  *Number of users using feature \( f_i \) around location \( l \in W \)*

- Estimation of \( \phi_i \) by **Non-parametric 2D-histogram estimator** on regular grid \( C \) of bandwidth \( w \)
  
  Small \( w \) \( \rightarrow \) Capture small scale variation/phenomena
  
  Large \( w \) \( \rightarrow \) Capture large scale variation/phenomena

\( ^1 \)motivated in next section
Problem Statement

- **Problem:**
  
  Given high-dimensional geographic feature signal $\Phi$ from a Social Media collection (all terms/tags)
  
  $\rightarrow$ Features might be meaningless, redundant, noisy

- **Goal:**
  
  Unsupervised extraction of small number of informative geographic features

- **Applications:**
  
  - Prepare data for learning tasks that cannot handle high-dimensional data
  - Discover hidden spatial variables in the data
Outline

1. Definitions and Problem Statement
   - Geographic Feature
   - Signal Estimation
   - Problem Statement

2. Data Characteristics and Normalization
   - Distribution Characteristics
   - Normalization
   - Geographic Feature Types

3. Latent Geographic Feature Extraction
   - Dimensionality Reduction
   - Framework

4. Experiments
   - Technique Comparison
   - Normalization Influence
   - Exploration Task

5. Conclusions
Dataset

- Two Flickr datasets covering US and LA
- Document features: Tags (pre-filtered by minimum user frequency)
- Spatial resolution: US (1.0 degree), LA (0.01 degree)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$\mathcal{D}_{US}$</th>
<th>$\mathcal{D}_{LA}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$</td>
<td>5,976,689</td>
<td>245,312</td>
</tr>
<tr>
<td>$</td>
<td>\mathcal{F}</td>
<td>$</td>
</tr>
<tr>
<td>$</td>
<td>U</td>
<td>$</td>
</tr>
<tr>
<td>$W$</td>
<td>US bounds (-124.87,25.25,-52.61,50.06)</td>
<td>LA city bounds (-124.87,25.25,-52.61,50.06)</td>
</tr>
<tr>
<td>$C_w$</td>
<td>$w = 1.0$ (111 km)</td>
<td>$w = 0.01$ (1.11 km)</td>
</tr>
<tr>
<td></td>
<td>$72 \times 24 = 1728$ cells</td>
<td>$43 \times 25 = 1075$ cells</td>
</tr>
<tr>
<td></td>
<td>585 cells with signal</td>
<td>945 cells with signal</td>
</tr>
</tbody>
</table>
Spatial Distribution Characteristics

**Figure:** $F(l)$: Num of features, $D(l)$: Num of documents, $U(l)$: Num of users, $Fd(l)$: Num of distinct features.

- Exponential characteristics of spatial feature distribution
- Users $\sim$ distinct features / documents $\sim$ features
Spatial Feature Distribution: 'beach'

Figure: \( F(l, f) \): Number of feature \( f = \text{beach} \), \( U(l, f) \): Number of users using \( f = \text{beach} \)

- Some users contribute large number of documents
- Estimate signal on basis of users is less biased (more robust)
Normalization

- Exponential distribution characteristics → Few locations dominate the signals’ spatial distribution
- Normalization transforms the signal into a more natural domain

Logging:
\[ \phi_i'(l) := \log \phi_i(l) + 1 \]

Binarization:
\[ \phi_i'(l) := 1\{\phi_i(l) > 0\} \]
**Geographic Feature Types**: Classes of geographic features with similar geographic semantics [Sengstock ACMGIS’11]

**Global**: Same intensity as baseline distribution (number of users) → *Not interesting to discriminate between locations*

**Regional**: Widely spread in geographic space but different from baseline → *Interesting to discriminate between large subsets in geographic space*

**Landmark**: Occurring only in small subsets of geographic space → *Interesting to discriminate between single small subset and the rest*

- Depends on area of interest $W$ and spatial resolution $w$. 
Geographic Feature Types

- **Entropy** over locations of spatial signal $X_i$ as geographic feature type statistic for $f_i$:

  
  large entropy $\rightarrow$ Signal widely spread / smoothly distributed  
  small entropy $\rightarrow$ Signal peaky / occurs in small areas

<table>
<thead>
<tr>
<th>Min</th>
<th>Median</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>massachusetts</td>
<td>sidewalk</td>
<td>nikon</td>
</tr>
<tr>
<td>seattle</td>
<td>waves</td>
<td>trees</td>
</tr>
<tr>
<td>losangeles</td>
<td>band</td>
<td>usa</td>
</tr>
<tr>
<td>sanfrancisco</td>
<td>smithsonian</td>
<td>landscape</td>
</tr>
<tr>
<td>chicago</td>
<td>olympus</td>
<td>nature</td>
</tr>
</tbody>
</table>

Figure: Ordered entropies $H[X_i]$ for tag features of US Flickr dataset
Outline

1. Definitions and Problem Statement
   - Geographic Feature
   - Signal Estimation
   - Problem Statement

2. Data Characteristics and Normalization
   - Distribution Characteristics
   - Normalization
   - Geographic Feature Types

3. Latent Geographic Feature Extraction
   - Dimensionality Reduction
   - Framework

4. Experiments
   - Technique Comparison
   - Normalization Influence
   - Exploration Task

5. Conclusions
Dimensionality Reduction

- Describe high-dimensional data by $k << p$ dimensions while preserving statistical properties of data.
- General formulation (generative latent factor model):

  $$X^{n\times p} = S^{n\times k} A^{k\times p}$$

  - $S$: Latent factor (component) values for each record
  - $A$: Combination of latent factors by original features

- **Latent Geographic Feature Sensor/Signal:**

  $$\tilde{\Phi}(l)^{k\times 1} := A^{k\times p} \Phi(l)^{p\times 1}$$
Spatial distributions in the data reflect spatial phenomena

Statistical structure in $X$ depends on spatial co-occurrence of features

Reducing the location sampling matrix $X$:

- Latent factors represent dominant spatial distributions
  
  (correlated features collapse, non-dominant features diminish)

- Latent factors describe distinct spatial distributions

Latent geographic features describe signals of dominant and distinct geographic phenomena
Framework

Input

1. Signal Estimation
   - Social Media \( D \)

2. Location Sampling
   - \( L, C_w \)

3. Dimensionality Reduction
   - Loc. Sampling Matrix \( X_{n \times p} \)
   - Reduced Loc. Sampling Matrix \( \tilde{X}_{n \times k} \)

Latent Geographic Feature Extraction

- \( w, C_w \)
- \( \Phi \)

Output

- Latent Geo. Feature Sensor \( \tilde{\Phi}, a \)
Outline

1. Definitions and Problem Statement
   - Geographic Feature
   - Signal Estimation
   - Problem Statement

2. Data Characteristics and Normalization
   - Distribution Characteristics
   - Normalization
   - Geographic Feature Types

3. Latent Geographic Feature Extraction
   - Dimensionality Reduction
   - Framework

4. Experiments
   - Technique Comparison
   - Normalization Influence
   - Exploration Task

5. Conclusions
Technique Comparison

- Study of dimensionality reduction techniques preserving different statistical properties
  - **Principal Component Analysis (PCA):** Components are statistically uncorrelated
  - **Sparse Principal Component Analysis (SPCA):** Components are statistically uncorrelated and sparse ($\alpha << p$ non-zero entries)
  - **Independent Component Analysis (ICA):** Components are statistically independent
Technique Comparison

- Qualitative Evaluation
  a. Extraction of $k = 20$ latent geographic features
  b. Manual labeling of extracted features on basis of component weights and spatial signal distribution
  c. Identification of 'informative' latent features
  d. Selection of similar latent features of other techniques on basis of highest component weights
  e. Comparison of component weights and signal distribution
**Technique Comparison**

**SPCA feat.:** 'landscape' (top), 'beach' (center), 'desert' (bot.)

- **landscape**: 0.95
- **mountain**: 0.24
- **national park**: 0.12
- **mountains**: 0.11
- **forest**: 0.07
- **sunset**: 0.03
- **clouds**: 0.03
- **trees**: 0.02

- **beach**: 0.76
- **ocean**: 0.61
- **sunset**: 0.18
- **boat**: 0.08
- **canon**: 0.06
- **people**: 0.06
- **white**: 0.05
- **sea**: 0.01

- **utah**: 0.63
- **arizona**: 0.54
- **desert**: 0.43
- **america**: 0.27
- **canyon**: 0.22
- **travel**: 0.09
- **rock**: 0.02
- **sky**: 0.00
Technique Comparison:

Comparison 'beach': PCA (top), ICA (center), SPCA (bottom)
Normalization Influence

- Extracted latent geographic features can be of different types (global, landmark, regional)

- Calculation of entropy over $k = 20$ extracted features for each technique and each normalization strategy (none, logging, binarization)

- SPCA and ICA show response towards more regional features for stronger normalization
Normalization Influence

- Normalization: None (top), Logging (center), Binar. (bottom)

- california 0.86
- sanfrancisco 0.32
- beach 0.07
- usa 0.01
- canada 0.00
- seattle 0.00
- losangeles -0.02
- newyork -0.39

- beach 0.76
- ocean 0.61
- sunset 0.18
- boat 0.08
- canon 0.06
- people 0.06
- white 0.05
- sea 0.01

- ocean 0.65
- beach 0.59
- sea 0.36
- boat 0.26
- sunset 0.16
- hdr 0.07
- longexposure 0.02
- america 0.00
Exploration Task

- Extraction of informative geographic features for Los Angeles using Flickr

- **Exploration Setting**
  - SPCA feature extraction
  - Normalization as exploration parameter
Exploration Task

Los Angeles Landmark Features

- ywoodboulevard 0.75
- walkofame 0.40
  - star 0.32
  - theater 0.30
- hollywood 0.24
- theatre 0.17
- street 0.04
- hotel 0.00
- littletokyo 0.82
  - train 0.47
  - station 0.26
- downtown 0.19
- metro 0.11
- street 0.04
- losangeles 0.02
- california 0.00
- boeing 0.58
- airport 0.48
  - lax 0.48
- airplane 0.32
- ernationalairport 0.23
- plane 0.22
- nikon 0.01
- 2010 0.00
Exploration Task

- Los Angeles Regional Features

- tourist 0.69
- d700 0.42
- lstatesofamerica 0.35
- d90 0.34
- vacation 0.26
- travel 0.15
- american 0.08
- america 0.06
- walk 0.74
- walking 0.59
- metro 0.21
- bicycle 0.18
- reetphotography 0.17
- sign 0.03
- bike 0.02
- cali 0.01
- flowers 0.83
- nature 0.35
- flower 0.24
- water 0.24
- bokeh 0.20
- light 0.15
- tree 0.10
- pink 0.07
Outline

1. Definitions and Problem Statement
   - Geographic Feature
   - Signal Estimation
   - Problem Statement

2. Data Characteristics and Normalization
   - Distribution Characteristics
   - Normalization
   - Geographic Feature Types

3. Latent Geographic Feature Extraction
   - Dimensionality Reduction
   - Framework

4. Experiments
   - Technique Comparison
   - Normalization Influence
   - Exploration Task

5. Conclusions
Conclusions

■ Summary
  ■ General framework for unsupervised extraction of informative spatial variables (dimensions of geographic semantics) from Social Media
  ■ PCA « ICA « SPCA
  ■ Transformation of informative geographic feature extraction into a problem of high-dimensional statistics in geographic feature space
  ■ Extraction of spatial variables of different types (landmarks, regional) by normalization

■ Ongoing Work
  ■ Extrinsic Evaluation of parametrization and techniques (e.g. spatial classification task)
  ■ (Semi-) supervised feature extraction