Multiple Location Profiling for Users and Relationships from Social Network and Content

Rui Li†  Shengjie Wang†  Kevin Chen-Chuan Chang*
† Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA
* Advanced Digital Sciences Center, Illinois at Singapore, Singapore
Outline

• Introduction
• Challenges
  – Location-based generation
  – Mixture of Observations
  – Partially available supervision
• Experiments
  – Home location prediction
  – Multiple location discovery
  – Relationship explanation
• Conclusion
Introduction

Users’ Locations are important for many information services

User

Content Provider

Local Content Recommendation

Local Friends Recommendation

Carol

Lives in: Los Angeles

and many others.
Introduction

• Many users of social networks are missing their location. Only 16% of users register their city and only 0.5% use GPS tags.

• People are likely to connect with people living nearby and tweet about close venues.

• Previous methods do not consider some issues.
Introduction

Carol in Real-world
Location: Los Angeles
Education: Uni. of Texas at Austin

The Following Network
- Bob
- Carol
- San Diego
- Mike
- Gaga
- Lucy
- Jean
- LA
- NY
- Austin

Tweets
- Terrible LA traffic!
- Want to go to Honolulu for Spring vacation!
- See Gaga in Hollywood.
- Good Morning!

Carol's Location Profile:
Los Angeles, Austin
Carol follows Lucy: Austin; Austin
Introduction

Informally:

The goal is to obtain the user’s multiple locations and profile both users and their relationship.
1) Location-based profiling
2) Mixture of observations
3) Partially available supervision
Location based generation

MLP models the probability that a relationship is based on the user’s location.

- Power-law distribution over distances for the following probability.

- Multinomial distribution over a set of venues for tweeting probability.
Location based generation

Power Law distribution

\[ p(x) \propto L(x)x^{-\alpha} \]

Sizes of earthquakes, craters on the moon, frequency of words in most languages, etc.
Location based generation

Multinomial Distribution

\[ f(x | n, p) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} \]

For \( n \) independent trials each of which leads to a success for exactly one of \( k \) categories, with given fixed success probabilities for each category.
Mixture of Observations

- Noisy signal challenge
  (Lady Gaga; Honolulu)
- Mixed signal challenge
  (Lucy and Bob)
Partially Available Supervision

- Only some users provide their locations. But this cannot be used as their profile.
- We cannot set the location assignments for his relationships as the observed location, as it does not allow the relationships to be generated based on other locations and fails to address the mixed-signal challenge.
More formally

**User and Relationship Location Profiling**

Given a set of users $U$, which contains both labeled users $U^*$ and unlabeled users $U^N$, the home location $l_{u_i}$ for $u_i \in U^*$, their following and tweeting relationships $f_{1:S}$ and $t_{1:K}$, and candidate locations $L$, estimate a set of locations $\hat{L}_{u_i} \subset L$ for $u_i \in U$, location assignments $\hat{x}_i \in \hat{L}_{u_i}$ and $\hat{y}_j \in \hat{L}_{u_j}$ for $f(i,j) \in f_{1:R}$, and a location assignment $\hat{z}_i \in \hat{L}_{u_i}$ for $t(i,j) \in t_{1:K}$, so as to make $\hat{L}_{u_i}$, $\hat{x}_i$, $\hat{y}_j$ and $\hat{z}_i$ close to $u_i$’s location profile $L_{u_i}$ and the true assignments $x_i$, $y_j$ and $z_i$ respectively.

$U=$\{Carol, Mike, Bob, Gaga, Lucy, Jean\}
$U^*=$\{Bob, Mike, Gaga, Lucy\}
$U^N=$\{Carol, Jean\}

$l_{Bob} =$ San Diego; $l_{Gaga} =$ NY; $l_{Lucy} =$ Austin; $l_{Mike} =$ LA
$f(Carol, Mike)=1; f(Carol, Bob)=1; f(Mike, Bob)=0$, etc..
$t(Carol, LA)=10; t(Carol, Honolulu)=1; t(Lucy, Austin)=5; t(Bob, Taj Mahal)=0$

$L_{Carol} =$ \{L.A., Austin\}

Given $f(Carol, Lucy)=1$, then $x_{Carol} = Austin$ $y_{Lucy} = Austin$ because they went to school together
Likewise, given $t(Carol, Hollywood)$ then $z_{Hollywood} = L.A.$ indicates that Carol is interested in $z_{Hollywood}
Multiple Location Profiling is a probabilistic generative approach, which models the joint probability of generating the two types of relationships based on users’ locations. We can estimate $\theta_i, x_i, y_j$ and $z_i$ with the observed relationships and locations, and use $\theta_i$ as $u_i$’s location profile.
1) generates each user $u_i$'s location distribution $\theta_i$, which is determined by the observed locations from the labeled users,

2) generates location assignments (e.g., $x_{s,i}$ and $z_{k,i}$) based on $\theta_i$, and

3) generates the associated following and tweeting relationships (e.g., $f_s\langle i,j \rangle$ and $t_k\langle i,j \rangle$) based on the location assignments.
Location based generation

1) the following probability decreases as the distance increases, and

2) at the distances in a long range, the probabilities do not decay as sharply as those at the distances in a short range.

(a) Following Probabilities versus Distances

\[ P(f(i,j)|\alpha, \beta, x_i, y_j) = \beta d(x_i, y_j)^\alpha \]
For each location, say Austin, we count the relative frequencies of the venues, and thus the probabilities, that the venues are tweeted by those users at the location.

1) nearby venues (e.g., “austin”) have high probabilities to be tweeted,
2) faraway venues (e.g., “hollywood”) have small probabilities to be tweeted, and
3) the probability to tweet a venue is not a monotonic function of its distance to the location.
Mixture of observations

• Noisy signal challenge

\( \mu \sim \text{Ber}(\rho_f) \) decides to use either location based model or random generative model. Parameter \( \rho_f \) models how likely a following relationship is generated.

The random following model \( F_R \) is also Bernoulli:

\[ P(f_{i,j} = 1 | F_R) = \frac{S}{N^2} \]

where \( S \) is the number of following relationships and \( N^2 \) is the total number of user pairs.
Mixture of observations

- Mixed Signal Challenge

Model a user $u_i$’s location profile as a multinomial distribution over candidate locations $L$, denoted as $\theta_i$. The probability of a location $l$ in $\theta_i$ represents how likely $u_i$ is at $l$. Our goal is to estimate $\theta_i$ for each $u_i$. We then assume that a location-based relationship is generated based on a specific location assignment picked from each related user’s profile, rather than their home locations only.

<table>
<thead>
<tr>
<th>Follower Location</th>
<th>Freq</th>
<th>Words</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Los Angeles</td>
<td>4</td>
<td>los angeles</td>
<td>1</td>
</tr>
<tr>
<td>Austin</td>
<td>3</td>
<td>paramount</td>
<td>1</td>
</tr>
<tr>
<td>San Diego</td>
<td>2</td>
<td>hollywood</td>
<td>2</td>
</tr>
<tr>
<td>Long Beach</td>
<td>1</td>
<td>austin</td>
<td>2</td>
</tr>
</tbody>
</table>

(c) Relationships as a Mixture of a User’s Locations
They chose to use the home locations of labeled users as prior knowledge to generate their location profiles.

\[ \Theta_i \sim \text{Dir}(\gamma) \]

\[ X \sim \text{Dir}(\alpha) \Rightarrow p(X) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha - 1} \]

\[ K \geq 2, \alpha = (\alpha_1, \alpha_2, ..., \alpha_K) \]

\[ \alpha_i > 0 \]
They use the available information to generate $\gamma$. The larger $\gamma$’s $l^{th}$ dimension $\gamma_l$ is, the more likely $\theta_i$ with a large probability in the $l$th dimension is to be generated.

Previous works set $\gamma$ as discrete uniform, but they use the information available:
They introduce an observation vector for each user $u_i$, denoted as $\eta_i \in \{0, 1\}^L$. Then, $\eta_{i,j}$ represents whether the jth location is observed. $\eta_{i,j} \sim Ber(b_o)$.

They introduce a boosting matrix $\Delta \in R^{L \times L}$, where $\Delta_{i,j}$ represents how much the prior of the jth location should be boosted when the ith location is observed. For simplicity, $\Delta = \text{diag}$, so the ith location only boosts its prior.

Thus, $\gamma_i = \eta_i \times \Delta \times \gamma + \gamma$ where the first term encodes how much we boost the prior for an observed location, and the second term encodes our priors for candidate locations. With $\gamma_i$, we will have a high probability to obtain $\theta_i$ that has a high probability to generate the observed location.
Experiments

- They used 3,980,061 user’s profiles and their social network.
- 630,187 Labeled users (cityName, stateName)
- 158,220 of these labeled users had at least one labeled friend or follower
- Only 139,180 user’s tweets were crawled, with at most 600 tweets per user.
- These were used as a data set. There are 14.8 friends, 14.9 followers, and 29.0 tweeted venues per user
Experiments

1) Home location prediction

Accuracy within 100 miles

\[ ACC@m = \frac{|\{u_i | u_i \in U \land d(l_u, \hat{l}_u) \leq m\}|}{|U|} \]

<table>
<thead>
<tr>
<th>Table 2: Home Location Prediction Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
</tr>
<tr>
<td>ACC@100</td>
</tr>
</tbody>
</table>
Experiments

A point \((x, y)\) means \(y\%\) of users were accurately predicted within \(x\) miles

(c) Overall Performance
2) Multiple Location Discovery

MANUALLY labeled 1000 users, and discovered 585 that clearly had multiple locations. On average, a user has 2 locations.

Distance-based Precision (DP)
Fraction of predicted locations that are “close enough” to true locations

\[
DP(u) = \frac{|\{l | l \in L'(u) \land c(l, L(u))\}|}{|L'(u)|}
\]

Distance-based Recall (DR)
Fraction of true locations that are close enough to predicted ones

\[
DR(u) = \frac{|\{l | l \in L(u) \land c(l, L'(u))\}|}{|L(u)|}
\]

Where \(c(l, L) = \text{True} \iff \exists l' \in L \text{ s.t. } D(l, l') < m\)
Experiments

### Table 3: Multiple Location Discovery Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Base(_U)</th>
<th>Base(_C)</th>
<th>MLP(_U)</th>
<th>MLP(_C)</th>
<th>MLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP@2</td>
<td>33.8%</td>
<td>39.3%</td>
<td>45.1%</td>
<td>48.3%</td>
<td>50.6%</td>
</tr>
<tr>
<td>DR@2</td>
<td>27.2%</td>
<td>33.1%</td>
<td>42.3%</td>
<td>45.3%</td>
<td>47.0%</td>
</tr>
</tbody>
</table>

![Figure 6: DP at Different Ranks](image1)

![Figure 7: DR at Different Ranks](image2)
3) Relationship Explanation

- They used the same 585 labeled users and kept the following relationships in which users’ location assignments could be clearly identified by their shared “regions”
- This generated 4,426 relationships and the location assignments of them.
- They used ACC@m and defined that a relationship is accurately explained if both users’ locations are accurately assigned within m miles.
- They compared against “Base”, which only considers the users’ home location and not the whole profile.
Experiments

Figure 8: Accuracy at Different Miles
Experiments

Table 5: Case Studies on Relationship Explanation

<table>
<thead>
<tr>
<th>User ID: 13069282, Location: Los Angeles</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Follower’s ID and</strong></td>
</tr>
<tr>
<td><strong>Follower’s Location</strong></td>
</tr>
<tr>
<td><strong>Location Assignments</strong></td>
</tr>
<tr>
<td>ID</td>
</tr>
<tr>
<td>--------------</td>
</tr>
<tr>
<td>101566144</td>
</tr>
<tr>
<td>14119630</td>
</tr>
<tr>
<td>15669188</td>
</tr>
<tr>
<td>53154473</td>
</tr>
</tbody>
</table>
Conclusion

• MLP is the first model that
  – Discovers users’ multiple locations
  – Profiles both user’s and relationships
• It is more accurate than previous methods in determining the user’s location
• It considers noisy and mixed data