Toward Mobile Cloud Computing: Data Analysis with Location-Based Social Network

Huan Liu

Joint Work with Huiji Gao and Jiliang Tang
Location-Based Social Networks (LBSNs)

- Location-Based Social Networking Sites
  - Foursquare, Facebook Places, Yelp

Real World

CHECK IN
Where the real world and OSNs meet

Online Social Networks
A Location-Based Social Network Framework

Social Computing

Content Layer
- Audios
- Videos
- Photos
- Tips

Social Layer

Geographical Layer

Timeline

Traditional Mobile Computing
Essential Data from LBSN

- Check-in history with time stamps
- Social networks derived from check-in locations
- User generated contents
- Interdependency of social networks and locations
Distinct Properties of LBSN Data

- Large-Scale Mobile Data
- Accurate Location Descriptions
- Explicit Social Friendships
- Significant Sparsity of Data
Research Opportunities

- Study a user’s mobile behavior through both **real** and **virtual worlds** in spatial, temporal and social dimensions.

- Understand the role of **social networks** and **geographical properties** with large amounts of heterogeneous data.

- Improve the development of **location-based services** such as mobile marketing, disaster relief, traffic forecasting, and etc.

- Mobile cloud computing
Some Challenges

- How to study human mobile behavior from high dimensional data from heterogeneous sources
- How to deduce human movement through sparse check-in data
- How to design location-based services to improve user’s experience without sacrificing one’s privacy
Potential Applications

- Disaster Relief/Crisis Response
- Mobile Search/Recommendation
- Location Prediction
- Recommendation Systems
- Mobile Community Detection
- Location Privacy Protection
- Mobile Marketing
Some of Our Recent Findings

• Social-Historical Ties on Location-Based Social Networks (ICWSM’2012)
  – Are two types of ties equally important?

• Geo-Social Correlation (CIKM’2012)
  – Handling the Cold Start Problem

• Mobile Location Prediction in Spatio-Temporal Context in Next Location Prediction in 2012 Nokia Mobile Data Challenge Workshop, 3rd Prize
  – Together is better
Social-Historical Effect of Online Check-ins
Why is the prediction hard

• Power-law distribution

(a) Power-law distribution of check-ins in whole dataset

Whole Dataset

(b) Power-law distribution of individual check-ins
Analyzing User’s Historical Ties

- **Short Term Effect**
  - The historical tie strength decreases over time.
  - The historical ties of the previous check-ins at airport, shuttle stop, hotel and restaurant have different strengths to the latest check-in of drinking coffee.
Modeling User’s Historical Ties

- Correspondences between language and LBSN modeling

<table>
<thead>
<tr>
<th>Language Modeling</th>
<th>LBSN Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td>Check-in collection</td>
</tr>
<tr>
<td>Document</td>
<td>Individual check-ins</td>
</tr>
</tbody>
</table>

|--------------------|-----------|----------|--------|------|--------------------|---------------------------|-------------------------|----------------------|------------------|

- Power-law distribution
- Short Term Effect

HPY (Hierarchical Pitman-Yor) Language Model
Modeling User’s Social Ties

- Social Ties
  - Common Check-ins

<table>
<thead>
<tr>
<th></th>
<th>Common check-ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>between friends</td>
<td>11.8306</td>
</tr>
<tr>
<td>between strangers</td>
<td>4.3226</td>
</tr>
</tbody>
</table>

- Check-in Similarities
  Users with friendship have higher check-in similarity than those without. Null hypothesis $H_0: S_{\mathcal{F}} \leq S_{\mathcal{R}}$, rejected at significant level $\alpha = 0.001$ with p-value of $2.6e-6$.
  - Friend Similarity
  - Friends’ Check-in Sequence
  - HPY

$$p_{SH}^i(c_{n+1} = l) = \eta P_H^i(c_{n+1} = l) + (1 - \eta) P_S^i(c_{n+1} = l)$$

Social Model
Experiment Results for Location Prediction

- Experiment Results
  - MFC
  - Most Frequent Check-in Model
  - MFT
  - Most Frequent Time Model
  - Order-1
  - Order-1 Markov Model
  - Order-2
  - Order-2 Markov Model
  - HM
  - Historical Model
  - SHM
  - Social-Historical Model
When no historical information is considered, the prediction performs worst, suggesting that considering social information only is not enough to capture the check-in behavior.

By gradually adding the historical information, the performance shows the following pattern: first increasing, reaching its peak value and then decreasing. Most of the time, the best performance is achieved at around $\eta = 0.7$. A big weight is given to historical ties, indicating that historical ties are more important than social ties.
Predicting New Check-Ins

Impossible to predict relying on personal history

Limited contribution to improve location prediction performance
Motivation

Geographical Distance

Social

Geo-Social Correlations

F: Local Friends : Local Non-friends
D: Distant Friends : Distant Non-friends
Geo-Social Correlations

\[ P_u^t(l) = \Phi_1 P_u^t(l|S_{FD}) + \Phi_2 P_u^t(l|S_{\neg FD}) + \Phi_3 P_u^t(l|S_{FD}) + \Phi_4 P_u^t(l|S_{\neg FD}). \]
$P^t_u(l)$: the probability of a user $u$ checking-in at a new location $l$ at time $t$

$$
P^t_u(l) = \Phi_1 P^t_u(l \mid S_{FD}) + \Phi_2 P^t_u(l \mid S_{FD}) \\
+ \Phi_3 P^t_u(l \mid S_{FD}) + \Phi_4 P^t_u(l \mid S_{FD}).
$$

Geo-Social Correlation Strength

$\Phi_1 = f(\mathbf{w}^T \mathbf{f}_u + b)$, $0 \leq \Phi_1 \leq 1$

$\Phi_2 = (1 - \Phi_1)\phi_1$

$\Phi_3 = (1 - \Phi_1)(1 - \phi_1)\phi_2$

$\Phi_4 = (1 - \Phi_1)(1 - \phi_1)(1 - \phi_2)$
Modeling Geo-Social Correlations

- $P_u^t(l)$: the probability of a user $u$ checking-in at a new location $l$ at time $t$

  $$
  P_u^t(l) = \Phi_1 P_u^t(l|S_{FD}) + \Phi_2 P_u^t(l|S_{FD}) + \Phi_3 P_u^t(l|S_{FD}) + \Phi_4 P_u^t(l|S_{FD}).
  $$

**Geo-Social Correlation Probability Measures:**

1. Sim-Location Frequency (S.Lf)
   $$
   P_u^t(l|S_x) = \frac{\sum_{v \in S_x} s(u,v) N_v^t(l)}{\sum_{v \in S_x} s(u,v) N_v^t}
   $$

2. Sim-User Frequency (S.Uf)
   $$
   P_u^t(l|S_x) = \frac{\sum_{v \in S_x} \delta_v^t(l) s(u,v)}{\sum_{v \in S_x} s(u,v)}
   $$

3. Sim-Location Frequency & User Frequency (S.Lf.Uf)
   $$
   P_u^t(l|S_x) = \frac{\sum_{v \in S_x} s(u,v) N_v^t(l)}{\sum_{v \in S_x} s(u,v) N_v^t} \frac{\sum_{v \in S_x} \delta_v^t(l)}{N_{S_x}}
   $$
Dataset

➢ Foursquare Dataset

Table 2: Statistical information of the dataset

<table>
<thead>
<tr>
<th>Duration</th>
<th>No. of user</th>
<th>No. of check-ins</th>
<th>No. of unique locations</th>
<th>No. of links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 1, 2011-July 31, 2011</td>
<td>11,326</td>
<td>1,385,223</td>
<td>182,968</td>
<td>47,164</td>
</tr>
</tbody>
</table>

Table 3: Statistical information of the July data

<table>
<thead>
<tr>
<th>Social Circle</th>
<th>No. of SCCs</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{FD}$</td>
<td>34,523</td>
<td>44.50%</td>
</tr>
<tr>
<td>$S_{FD}$</td>
<td>5,636</td>
<td>7.26%</td>
</tr>
<tr>
<td>$S_{FD}$</td>
<td>3,588</td>
<td>4.62%</td>
</tr>
<tr>
<td>$S_{FD}$</td>
<td>39,423</td>
<td>50.82%</td>
</tr>
<tr>
<td>Others</td>
<td>1,672</td>
<td>2.2%</td>
</tr>
<tr>
<td>$S_{FD} \cup S_{FD}$</td>
<td>35,277</td>
<td>45.47%</td>
</tr>
<tr>
<td>$S_{FD} \cup S_{FD}$</td>
<td>35,784</td>
<td>46.12%</td>
</tr>
<tr>
<td>$S_{FD} \cup S_{FD}$</td>
<td>8,235</td>
<td>10.61%</td>
</tr>
<tr>
<td>$S_{FD} \cup S_{FD} \cup S_{FD}$</td>
<td>36,486</td>
<td>47.03%</td>
</tr>
</tbody>
</table>
Experiments

- Location Prediction Evaluation Metrics

<table>
<thead>
<tr>
<th></th>
<th>Single Measure</th>
<th>Various Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal Strength</td>
<td>EsSm</td>
<td>EsVm</td>
</tr>
<tr>
<td>Random Strength</td>
<td>RsSm</td>
<td>RsVm</td>
</tr>
<tr>
<td>Various Strength</td>
<td>VsSm</td>
<td>gSCorr</td>
</tr>
</tbody>
</table>

- Effect of Geo-Social Correlation Strength and Probability Measures

<table>
<thead>
<tr>
<th>Methods</th>
<th>Top-1</th>
<th>Top-2</th>
<th>Top-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>EsVm</td>
<td>17.88%</td>
<td>24.06%</td>
<td>27.86%</td>
</tr>
<tr>
<td>EsSm</td>
<td>16.20%</td>
<td>21.92%</td>
<td>25.43%</td>
</tr>
<tr>
<td>VsSm</td>
<td>16.49%</td>
<td>22.28%</td>
<td>25.92%</td>
</tr>
<tr>
<td>RsSm</td>
<td>14.93%</td>
<td>20.30%</td>
<td>23.70%</td>
</tr>
<tr>
<td>RsVm</td>
<td>15.23%</td>
<td>20.85%</td>
<td>24.50%</td>
</tr>
<tr>
<td>gSCorr</td>
<td>19.21%</td>
<td>25.19%</td>
<td>28.69%</td>
</tr>
</tbody>
</table>
### Experiments

#### Effect of Different Geo-Social Circles

<table>
<thead>
<tr>
<th>Methods</th>
<th>Top-1</th>
<th>Top-2</th>
<th>Top-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.51%</td>
<td>8.31%</td>
<td>9.32%</td>
<td></td>
</tr>
<tr>
<td>3.65%</td>
<td>4.75%</td>
<td>5.34%</td>
<td></td>
</tr>
<tr>
<td>18.37%</td>
<td>24.10%</td>
<td>27.34%</td>
<td></td>
</tr>
<tr>
<td>18.62%</td>
<td>24.44%</td>
<td>27.79%</td>
<td></td>
</tr>
<tr>
<td>19.01%</td>
<td>24.95%</td>
<td>28.35%</td>
<td></td>
</tr>
<tr>
<td>8.33%</td>
<td>10.79%</td>
<td>12.23%</td>
<td></td>
</tr>
<tr>
<td><strong>19.21%</strong></td>
<td><strong>25.19%</strong></td>
<td><strong>28.69%</strong></td>
<td></td>
</tr>
</tbody>
</table>
Mobile Location Prediction in Spatio-Temporal Context

LBSN

Temporal

Spatial

Content

Social

Mobile Location Prediction in Spatio-Temporal Context
Problem Statement

The probability of checking in at location \( l \) given the check-in time at \( t \) and latest check-in:

\[
p(v_i = l \mid t_i = t, v_{i-1} = l_k) = p(t_i = t \mid v_i = l)p(v_i = l \mid v_{i-1} = l_k)
\]

Temporal Constraint

The probability of the \( i \)-th visit happening at time \( t \), observing that the \( i \)-th visit location is \( l \).

Spatial Prior

The probability of next visit at location \( l \) given the current visit at \( l_k \).

Historical Model
Temporal Constraint:

\[ p(t_i = t | v_i = l) \]

\[ = p(h_i = h, d_i = d | v_i = l) \]

\[ = p(h_i = h | v_i = l) p(d_i = d | v_i = l) \]

**Hourly Constraint**  **Daily Constraint**

h: Hour of the day, i.e., 10:00am, 3:00pm

d: Day of the week, i.e., Monday, Sunday
Temporal Constraint

Compute $p(h_i = h \mid v_i = l)$ and $p(d_i = d \mid v_i = l)$

- Distribution of a user’s visits at a specific location in 24 hours.
  (user id: 013; place id: 3)

\[
p(h_i = h \mid v_i = l) = N_l(h \mid \mu_h, \sigma_h^2)
\]

\[
p(H \mid v_i = l) = \prod_{i=1}^{N_l} N_l(h_i \mid \mu_h, \sigma_h^2)
\]

\[
(h_i \in H, \mid H \mid = N_l)
\]

Maximizing Likelihood

\[
\{ \mu_h, \sigma_h^2 \} 
\]
Temporal Constraint

Curve Fitting:

[user id: 013; place id: 3]
Probability of visiting location $l$ at time $t$ with the latest visit at $l_k$

\[
p(v_i = l \mid t_i = t, v_{i-1} = l_k)
= p(v_i = l \mid v_{i-1} = l_k)p(h_i = h \mid v_i = l)p(d_i = d \mid v_i = l)
= p(v_i = l \mid v_{i-1} = l_k)N_l(h \mid \mu_h, \sigma_h^2)N_l(d \mid \mu_d, \sigma_d^2)
\]

HPY Prior Gaussian Gaussian

HPY Prior Hour-Day Model (HPHD)
Experiments – Together is Better

**Results**

<table>
<thead>
<tr>
<th>Models</th>
<th>Correct No.</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial-based</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFV</td>
<td>1148</td>
<td>0.3402</td>
</tr>
<tr>
<td>OMM</td>
<td>1466</td>
<td>0.4345</td>
</tr>
<tr>
<td>FMM</td>
<td>1583</td>
<td>0.4692</td>
</tr>
<tr>
<td>HP</td>
<td>1610</td>
<td>0.4772</td>
</tr>
<tr>
<td>Temporal-based</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFH</td>
<td>1462</td>
<td>0.4333</td>
</tr>
<tr>
<td>MFD</td>
<td>1156</td>
<td>0.3426</td>
</tr>
<tr>
<td>MFHD</td>
<td>1538</td>
<td>0.4558</td>
</tr>
<tr>
<td>Spatio-temporal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPH</td>
<td>1680</td>
<td>0.4979</td>
</tr>
<tr>
<td>HPD</td>
<td>1583</td>
<td>0.4692</td>
</tr>
<tr>
<td><strong>HPHD</strong></td>
<td><strong>1705</strong></td>
<td><strong>0.5053</strong></td>
</tr>
</tbody>
</table>

Rank 3rd among 21 participated teams in Nokia Mobile Competition
Some of Our Recent Findings

• Social-Historical Ties on Location-Based Social Networks (ICWSM’2012)
  – Are two types of ties equally important?
• Geo-Social Correlation (CIKM’2012)
  – Handling the Cold Start Problem
• Mobile Location Prediction in Spatio-Temporal Context in Next Location Prediction in 2012 Nokia Mobile Data Challenge Workshop, 3rd Prize
  – Together is better
Acknowledgments: The projects are, in part, sponsored by ONR grants.

THANK YOU