Exploiting Spatiotemporal User Behaviours for User Linkage

Wei Chen¹, Hongzhi Yin², Weiqing Wang²
Lei Zhao¹, Wen Hua², Xiaofang Zhou²

¹ School of Computer Science and Technology, Soochow University, China
³ School of ITEE, The University of Queensland, Brisbane, Australia
Outline

• Introduction
• Problem Statement
• Feature Extraction
• Similarity Measure
• Experiments
• Conclusion
Introduction

• The proliferation of GPS-enabled devices and mobile techniques has led to the emergence of large amount of spatiotemporal information.
  – Trajectory data: adjacent points of a trajectory are sampled in a short time period.
  – Discrete check-in data in social network: the time between two check-ins is usually large.

![Trajectory](image1.png)  ![Discrete check-in record](image2.png)

Trjectory  Discrete check-in record
Introduction

• Spatiotemporal data based studies:
  – Route planning in road networks
  – Activity trajectory recommendation
  – Understand human mobility pattern
  – ……
  – Cross-domain user linkage with spatiotemporal data [1]
Introduction

• Cross-domain user linkage: link the same user across different domains

  Domain A  Domain B

  Example:  Facebook---Twitter
Problem Statement

- Spatiotemporal record
  - A spatiotemporal record on both trajectory data and check-in data is defined as: \( d = (u, lat, lng, t) \)
  - \( u \): the unique id of a user
  - \( lat \): latitude of the record
  - \( lng \): longitude of the record
  - \( t \): timestamp of the record

- Example

| adnys, 34.0553261066, 118.246986866, 07:50:22 | 1808.42MB | 02:06:59 |
| adnys, 34.0314662928, 118.462771922, 08:15:33 | 1808.42MB | 02:06:59 |
| adnys, 38.5494890477, 121.740014302, 12:29:09 | 1808.42MB | 02:06:59 |
| adnys, 33.1258538651, 117.311940422, 15:41:15 | 1808.42MB | 02:06:59 |
| adnys, 34.259089, 116.867585778, 06:27:12 | 1808.42MB | 02:06:59 |
| adnys, 34.1838325292, 118.275179542, 11:34:44 | 1808.42MB | 02:06:59 |
| adnys, 21.2819466313, 157.836713791, 16:26:08 | 1808.42MB | 02:06:59 |
| adnys, 33.9847729546, 118.449375629, 20:11:25 | 1808.42MB | 02:06:59 |
| adnys, 37.7896414689, 122.394288182, 21:43:17 | 1808.42MB | 02:06:59 |
Problem Statement

- Two kinds of important data
  - Check-in data, which can be used to extract features directly.
  - Trajectory data, which needs preprocessing before extracting features.
Problem Statement

- **Stay point** [2]: a stay point $s$ stands for a geographic region where a user stayed over a certain time interval.

  - Given a trajectory $\tau = (p_1, p_2, \ldots, p_n)$, if there exists a group of consecutive points $P = (p_i, p_{i+1}, \ldots, p_j)$ of $\tau$ such that $\forall i < k \leq j$, $Distance(p_i, p_k) \leq \delta_d$ and $|p_j\cdot t - p_k\cdot t| \geq \delta_d$ then we have a stay point $s$ in the form of

    $$(s.\text{lat}, s.\text{lng}) = \left(\frac{\sum_{k=i}^{j} p_k\cdot \text{lat}}{|P|}, \frac{\sum_{k=i}^{j} p_k\cdot \text{lng}}{|P|}\right)$$

Problem Statement

- **Stay region candidate point**
  - Given a trajectory $\tau = (p_1, p_2, \ldots, p_n)$, the start point $p_1$, the end point $p_n$, each point of $P$ is defined as stay region candidate point, denoted as $r_c$.

- **Example**

![Trajectory Image]
Problem Statement

- Semantics behind the check-ins and stay region candidate points:
  - Shopping mall
  - Home region
  - Work region
  - Bus station
  - ……
Problem Statement

• Formulation: Given user sets $U_1 = \{u_{11}, u_{12}, \ldots, u_{1n}\}$ and $U_2 = \{u_{21}, u_{22}, \ldots, u_{2m}\}$, where each user is associated with a set of spatiotemporal records, we aim at finding linked user pairs across these two domains.
User Linkage

• Extract features
• Measure user similarity
Feature Extraction

• Features
  – Stay region distribution
  – Global time distribution
  – Local time distribution
Feature Extraction

• Stay region distribution [3]

\[ p = \sum_j \chi \left( d_{r_c^i, r_c^j} - d_c \right), \quad \begin{cases} \chi(x) = 1, & \text{if } x < 0 \\ \chi(x) = 0, & \text{otherwise} \end{cases} \]

\[ \delta = \begin{cases} \min_{p_{r_c^j} > p_{r_c^i}} (d_{r_c^i, r_c^j}), & \text{if } p_{r_c^j} > p_{r_c^i} \\ \max_j (d_{r_c^i, r_c^j}), & \text{otherwise} \end{cases} \]

Feature Extraction

• Example
Feature Extraction

• Region weight calculation.
  – In real life, many people tend to visit popular areas, such as the downtown of a city, a large bus station, and a popular cinema. Obviously, the importance of the extracted stay regions are diverse.
  – Highlight the individual region.
  – Lighten the popular region.
Feature Extraction

- Region weight calculation.

(a) User Region

<table>
<thead>
<tr>
<th>User</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>$(R_1^1, \cdots, R_1^l)$</td>
</tr>
<tr>
<td>$u_2$</td>
<td>$(R_2^1, \cdots, R_2^k)$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$u_n$</td>
<td>$(R_n^1, \cdots, R_n^m)$</td>
</tr>
</tbody>
</table>

(b) Region Weight

<table>
<thead>
<tr>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>${\omega(R_1^1), \cdots, \omega(R_1^l)}$</td>
</tr>
<tr>
<td>${\omega(R_2^1), \cdots, \omega(R_2^k)}$</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>${\omega(R_n^1), \cdots, \omega(R_n^m)}$</td>
</tr>
</tbody>
</table>

$$\omega(R_i^j) = \frac{N}{1 + \sum S(R_i^j, R_o)}$$
Feature Extraction

• Example

• Note: the points outside the region are omitted.
Feature Extraction

• Spatiotemporal features
  – Stay region distribution
  – Global time distribution
  – Local time distribution
Feature Extraction

• Global time distribution
  – We extract the temporal features from the global perspective, where the stay region factor is omitted.
  – Given a set of stay region candidate points \((r_c^1, r_c^2, \ldots, r_c^n)\) of a user \(u\), the Expectation Maximization (EM) algorithm is used to find optimal parameters with timestamp set \((r_c^1.t, r_c^2.t, \ldots, r_c^n.t)\).

• Example

<table>
<thead>
<tr>
<th>adnys, 34.0553261066, 118.246986866, 07:50:22</th>
<th>adnys, 34.0314662928, 118.462771922, 08:15:33</th>
</tr>
</thead>
<tbody>
<tr>
<td>adnys, 38.5494890477, 121.740014302, 12:29:09</td>
<td>adnys, 33.1258538651, 117.311940422, 15:41:15</td>
</tr>
<tr>
<td>adnys, 34.259089, 116.867585778, 06:27:12</td>
<td>adnys, 34.1838325292, 118.275179542, 11:34:44</td>
</tr>
<tr>
<td>adnys, 37.7896414689, 122.394288182, 21:43:17</td>
<td></td>
</tr>
</tbody>
</table>
Feature Extraction

- Global time distribution
  - E-step: the probability of the sample $r^i_c.t$ generated by the cluster $(\mu_k, \Sigma_k)$ is:
    \[
    \gamma_{ik} = \frac{\alpha_k N(r^i_c.t | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \alpha_k N(r^i_c.t | \mu_j, \Sigma_k)}
    \]
  - M-step: the maximum likelihood method is used to update model parameters as follows:
    \[
    \alpha_k = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ik}
    \]
    \[
    \mu_k = \frac{\sum_{i=1}^{n} \gamma_{ik} r^i_c.t}{\sum_{i=1}^{n} \gamma_{ik}}
    \]
    \[
    \Sigma_k = \frac{\sum_{i=1}^{n} \gamma_{ik} (r^i_c.t - \mu_k)^2}{\sum_{i=1}^{n} \gamma_{ik}}
    \]
Feature Extraction

- Global time distribution

(a) Time Cluster

<table>
<thead>
<tr>
<th>User</th>
<th>Time Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>$(T_1^1, \cdots, T_1^i)$</td>
</tr>
<tr>
<td>$u_2$</td>
<td>$(T_2^1, \cdots, T_2^k)$</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>$u_n$</td>
<td>$(T_n^1, \cdots, T_n^m)$</td>
</tr>
</tbody>
</table>

(b) Time Cluster Weight

<table>
<thead>
<tr>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>${\omega(T_1^1), \cdots, \omega(T_1^i)}$</td>
</tr>
<tr>
<td>${\omega(T_2^1), \cdots, \omega(T_2^k)}$</td>
</tr>
<tr>
<td>\ldots</td>
</tr>
<tr>
<td>${\omega(T_n^1), \cdots, \omega(T_n^m)}$</td>
</tr>
</tbody>
</table>

$$
\omega(T_1^i) = \frac{\sum_{i=1}^{N} S(T_1^i, T_o) \bigg / \sum_{i=1}^{N} 1 + \sum_{i=1}^{N} S(T_1^i, T_o)}
$$
Feature Extraction

• Spatiotemporal features
  – Stay region distribution
  – Global time distribution
  – Local time distribution
Feature Extraction

• Local time distribution
  – We use the same method to extract time clusters and calculate corresponding weights in each stay region.

• Example
User Linkage

- Extract feature
- Measure similarity
Similarity Measure

- Stay region similarity
  - Assume the stay regions of $u_1$ and $u_2$ are:
    $$\{(R_1^1, \omega(R_1^1)), (R_1^2, \omega(R_1^2)), \ldots, (R_1^m, \omega(R_1^m))\}$$
    $$\{(R_2^1, \omega(R_2^1)), (R_2^2, \omega(R_2^2)), \ldots, (R_2^n, \omega(R_2^n))\}$$
  - The stay region similarity $S(u_1, u_2)_r$ is defined as:
    $$S(u_1, u_2)_r = \sum_{i=1}^{m} \sum_{j=1}^{n} S(R_1^i, R_2^j) \omega(R_1^i) \omega(R_2^j)$$
Similarity Measure

- Global time similarity.
  - Assume the global time clusters of \( u_1 \) and \( u_2 \) are:
    \[
    \{(T_1^1, \omega(T_1^1)), (T_1^2, \omega(T_1^2)) \ldots, (T_1^k, \omega(T_1^k))\}
    \]
    \[
    \{(T_2^1, \omega(T_2^1)), (T_2^2, \omega(T_2^2)) \ldots, (T_2^l, \omega(T_2^l))\}
    \]
  - The global time similarity \( S(u_1, u_2)_t \) is defined as:
    \[
    S(u_1, u_2)_t = \sum_{i=1}^{k} \sum_{j=1}^{l} S(T_1^i, T_2^j) \omega(T_1^i) \omega(T_2^j)
    \]
Similarity Measure

- Local time similarity.
  - Assume the time distribution in a stay region \((R_1^i, \omega(R_1^i))\) of \(u_1\) is \(((T_1^1, \omega(T_1^1)), (T_1^2, \omega(T_1^2)), \ldots, (T_1^k, \omega(T_1^k)))\), in a stay region \((R_2^j, \omega(R_2^j))\) of \(u_2\) is \(((T_2^1, \omega(T_2^1)), (T_2^2, \omega(T_2^2)), \ldots, (T_2^l, \omega(T_2^l)))\), the local time similarity in these two regions is defined as:
    \[
    S(R_1^i, R_2^j) \omega(R_1^i) \omega(R_2^j) \sum_{i=1}^{k} \sum_{j=1}^{l} S(T_1^i, T_2^j) \omega(T_1^i) \omega(T_2^j)
    \]
  - The local time similarity between \(u_1\) and \(u_2\) is defined as:
    \[
    S(u_1, u_2)_{rt} = \sum_{i=1}^{m} \sum_{j=1}^{n} (S(R_1^i, R_2^j) \omega(R_1^i) \omega(R_2^j) \sum_{i=1}^{k} \sum_{j=1}^{l} S(T_1^i, T_2^j) \omega(T_1^i) \omega(T_2^j))
    \]
Similarity Measure

- Finally, the similarity between \( u_1 \) and \( u_2 \) is defined as:

\[
S(u_1, u_2) = S(u_1, u_2)_r + S(u_1, u_2)_t + S(u_1, u_2)_{rt}
\]
Experiments

• Dataset
  – Beijing Walk Trajectories (BJW) -- Beijing Car Trajectories (BJC)
  – Foursquare (FS) -- Twitter (TW)
  – Instagram (IT) -- Twitter (TW)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Domain</th>
<th>Users</th>
<th>Trajectories</th>
<th>Locations/Check-ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>BJW-BJC</td>
<td>Walk</td>
<td>182</td>
<td>14337</td>
<td>2190957</td>
</tr>
<tr>
<td></td>
<td>Car</td>
<td>182</td>
<td>5475</td>
<td>925380</td>
</tr>
<tr>
<td>FS-TW</td>
<td>Foursquare</td>
<td>282</td>
<td>-</td>
<td>7832</td>
</tr>
<tr>
<td></td>
<td>Twitter</td>
<td>282</td>
<td>-</td>
<td>88820</td>
</tr>
<tr>
<td>IT-TW</td>
<td>Instagram</td>
<td>1066</td>
<td>-</td>
<td>283740</td>
</tr>
<tr>
<td></td>
<td>Twitter</td>
<td>1066</td>
<td>-</td>
<td>284051</td>
</tr>
</tbody>
</table>
Experiments

• Compared methods:
  – **GC**: Each user is denoted by a set of grid cells.
  – **LT**: Each user is presented by a set of bins.
  – **STUL-S**: A simplified version of STUL, where the extracted features are directly used to measure the user similarity.

• Our approach:
  – **STUL**
Experiments

• Evaluation metrics:
  
  - $precision = \frac{k}{n}$
  
  - $recall = \frac{k}{m}$
  
  - $F1 = \frac{2 \times Recall \times Precision}{Recall + Precision}$
Experiments

• Performance of the proposed algorithms in different datasets
Experiments

- Performance of STUL w.r.t varied $\theta$
Experiments

- Performance of STUL w.r.t. varied cutoff distance

![Graphs showing performance metrics of STUL with varied cutoff distance for different methods: (a) BJW-BJC, (b) FS-TW, (c) IT-TW.](image)
Experiments

- Performance of STUL w.r.t. varied $\xi$
Conclusion

• To connect the actually linked users from different domains with spatiotemporal data, we propose the novel model STUL.
  – From spatial perspective, a density-based method is developed to extract stay regions that a user will visit repeatedly.
  – From temporal perspective, we use GMM to extract the time distribution. Based on these features, we measure the similarity between users. The real-world dataset based experiments demonstrate the high performance of STUL.
Thank You
Q & A