Embedding-based Mobility Pattern Mining

Kangyan Zhou (kzhou10)
Lunan Li (lunanli3)
Hong Cheng (hcheng17)
Why Embedding and Mobility Pattern?

- **Motivation:** User group specific Mobility Patterns
- **Possible usage:** people flow prediction
  - Traffic scheduling
  - Peer recommendation
- **Our work is based on Regions, Periods, Activities [1]**
  - Embedding for activities discovery; no user involvement
- **Other previous work on mobility pattern**
  - An Empirical Study of Geographic User Activity Patterns in Foursquare [3]
  - Spatial information only: no semantic
Why Embedding and Mobility Pattern?

- **Frequent Triplet**
  - Weekday, hour, hotspots
  - Intuition: people tend to conduct certain event at a specific time
  - E.g. go to club at night on the weekend

- **Trajectory**
  - Transitions among hotspots
  - Impose time and location constraints
  - E.g: from Siebel center to ARC

- **Result Demo**
<table>
<thead>
<tr>
<th>#</th>
<th>Keyword</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>restaurant</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>cocina</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>terranea</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>hospitality</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>richmond</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>euphoria</td>
<td>22</td>
</tr>
<tr>
<td>7</td>
<td>ranchpalosverdes</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>malo</td>
<td>20</td>
</tr>
<tr>
<td>9</td>
<td>sweet rose</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>restauration</td>
<td>11</td>
</tr>
</tbody>
</table>
### User Group

**User Group 1**

#### User Group Information

<table>
<thead>
<tr>
<th>Hotspot Number</th>
<th>Center</th>
<th>Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>34.05369929113924, -118.24978725632911</td>
<td>[&quot;dodger&quot;, &quot;stadium&quot;, &quot;dtla&quot;, &quot;lakers&quot;, &quot;downtown&quot;]</td>
</tr>
<tr>
<td>11</td>
<td>34.101012796875, -118.333811890625</td>
<td>[&quot;hollywood&quot;, &quot;bowl&quot;, &quot;nightclub&quot;, &quot;walk&quot;, &quot;fame&quot;]</td>
</tr>
<tr>
<td>14</td>
<td>34.08227549350649, -118.37248937662336</td>
<td>[&quot;west&quot;, &quot;alfred&quot;, &quot;melrose&quot;, &quot;left&quot;, &quot;marmont&quot;]</td>
</tr>
</tbody>
</table>

#### Sequential Patterns

#### Frequent Triplets

---

**Map**

- Simi Valley
- Thousand Oaks
- Santa Monica Mountains National Recreation Area
- Topanga State Park
- Malibu
- Beverly Hills
- Los Angeles

**Satellite**

- Santa Monica Mountains
- Point Mugu State Park
- Azusa
- City of Industry
- Covina
- West Covina
- Pomona

---

**Terms of Use**

- © 2017 Google

---

**Report a map error**

- Chino Hills

---

**Google Maps**
### Hot Spots

<table>
<thead>
<tr>
<th></th>
<th>Count Percent</th>
<th>Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>[3,11]</td>
<td>0.0268</td>
<td>stadium-&gt;night club</td>
</tr>
<tr>
<td>[11,14]</td>
<td>0.0514</td>
<td>night club-&gt;hotel</td>
</tr>
<tr>
<td>[14,11]</td>
<td>0.0491</td>
<td>hotel-&gt;night club</td>
</tr>
</tbody>
</table>
### User Group Information

<table>
<thead>
<tr>
<th>Triplet</th>
<th>Count Percent</th>
<th>Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sat, 2, 11</td>
<td>0.0031</td>
<td>[&quot;hollywood&quot;, &quot;whiskeybluhw&quot;, &quot;joeybuicesound&quot;, &quot;supperclub&quot;, &quot;whiskey blu&quot;]</td>
</tr>
<tr>
<td>Sat, 1, 11</td>
<td>0.0026</td>
<td>[&quot;hollywoodbowl&quot;, &quot;hollywood&quot;, &quot;tonight&quot;, &quot;whiskeybluhw&quot;, &quot;davestewart&quot;]</td>
</tr>
<tr>
<td>Sat, 21, 11</td>
<td>0.0022</td>
<td>[&quot;hollywood&quot;, &quot;night&quot;, &quot;event&quot;, &quot;find&quot;, &quot;game&quot;]</td>
</tr>
</tbody>
</table>
Outline

● Previous Work: Regions, Periods, Activities
● System Overview
● System Details
  ○ Data Preprocessing
  ○ Geo-tagged Data Embedding
  ○ User Grouping
  ○ Hotspot Detection
  ○ Mobility Pattern Discovery
● Future Work
Previous Work: Regions, Periods, Activities

- **Input:** geo-tagged Data (tweets)
  - Tweet text
  - Longitude, latitude
  - Time
  - User id
- **Algorithm:** build a heterogeneous network embedding with unigram(word), location and time
- **Output:** a query-based system that takes any one or two of the three elements as input, and output the top hits
System Overview

**DATA/EMBED**

- Tweets
  - Data Preprocess: Filter advertisements, delete repeated tweets
  - Embedding: User/Word/Location/Time
  - User Grouping
  - User Embedding
  - Word Embedding

**HOTSPOT/MOBILITY PATTERN**

- Trajectory Pattern
- Frequent Triplets
- Time/Location
- Location
- Mobility Pattern Discovery
- Hotspots Detection
Data Preprocessing

● Current Issue
  ○ Advertisements
    ■ Content of advertisements are not related to the location
    ■ Advertisements share many similar keywords that affect our semantic meaning
  ○ Repeated Tweets
    ■ Same tweets are posted multiple times by twitter bots
    ■ Also affect semantic meaning and mobility pattern discovery

● Approach
  ○ Set Approach
  ○ User Average Similarity
Data Preprocessing

- **Set Approach**
  - For every user $u$
    - $n$ - the number of original tweets
    - Put all original tweets in a Set $K$
    - $s$ - the size of $K$
    - Shrink Ratio = $\frac{s}{n}$

- Remove 185 users
Data Preprocessing

- Set Approach Result
Data Preprocessing

- Set Approach Result

```
501049977817563136 SOH 907238670 SOH 34.1718815SOH-118.5891375SOH Sun Aug 17 11:56:38 CDT 2014 SOH 461588198 SOH using mastercard lot hope
   get pricelesssurprises jtimmerlake related please pick tix SOH I've been using my @MasterCard a lot in hopes to get some
#PricelessSurprises that are @jtimmerlake related! Please pick me for some tix! SOH SOH SOH SOH SOH

501050013926318080 SOH 907238670 SOH 34.171875SOH-118.5892025SOH Sun Aug 17 11:56:47 CDT 2014 SOH 461588207 SOH using mastercard lot hope
   get pricelesssurprises jtimmerlake related please pick tix SOH I've been using my @MasterCard a lot in hopes to get some
#PricelessSurprises that are @jtimmerlake related! Please pick me for some tix! 😄 SOH SOH SOH SOH SOH

501050049309462530 SOH 907238670 SOH 34.1719355SOH-118.5889615SOH Sun Aug 17 11:56:55 CDT 2014 SOH 461588215 SOH using mastercard lot hope
   get pricelesssurprises jtimmerlake related please pick tix SOH I've been using my @MasterCard a lot in hopes to get some
#PricelessSurprises that are @jtimmerlake related! Please pick me for some tix! 😄😄 SOH SOH SOH SOH SOH

501050067172982784 SOH 907238670 SOH 34.1719355SOH-118.5889615SOH Sun Aug 17 11:56:59 CDT 2014 SOH 461588219 SOH using mastercard lot hope
   get pricelesssurprises jtimmerlake related please pick tix SOH I've been using my @MasterCard a lot in hopes to get some
#PricelessSurprises that are @jtimmerlake related! Please pick me for some tix! 😄😄😄 SOH SOH SOH SOH SOH
```
Data Preprocessing

● User Average Similarity
  ○ For every user $u$
    ■ Compute the similarities of every two tweets
    ■ Similarity score ranges from 0 to 1:
      ● cosine similarity (word embedding)
      ● Python SequenceMatcher (ratio of same words in two tweets)
    ■ Adds up all the similarities and compute the average
  ○ Delete the user if the average cosine similarity is larger or equal than 0.8
Data Preprocessing

- User Average Similarity Result
Embedding in Heterogeneous Information Network

- Only select users that send over 45 tweets
  - In average around 3 tweets per day -- suitable for our task
- Three types of node already in the graph: location, word, time
  - Location: use the mean-shift algorithm to group nearby coordinates
  - Time: map time stamp into seconds. 8:00:00 am = 8*60*60s
  - Word: only select the most 20000 words.
- Types of edges: location-word, word-word, location-time, and word-time
  - Weighted edges
  - Word-word co-occurrence in each tweet
- Include user in the heterogeneous information network by adding a node type ‘user’ and the corresponding edges: user-location, user-time, user-word
Embedding in Heterogeneous Information Network

- **Training Objective**
  - Optimize the sum of losses between different types of subnetworks.
  - Use second-order proximity(1) in LINE to measure node similarity.

- **Training Process**
  - Use negative sampling(2) to generate node embedding for users and words.

- **Output word embedding and user embedding for next round of use**

\[
p_1(v_j | v_i) = \frac{\exp(u_j^T \cdot \bar{u}_i)}{\sum_{k=1}^{\left|V\right|} \exp(u_k^T \cdot \bar{u}_i)} \tag{1}
\]

\[
L = -\log\sigma(v_j^T \cdot v_i) - \sum_{K=1}^{k} \log\sigma(v_k^T \cdot v_i) \tag{2}
\]
User Grouping

- Hierarchical Clustering
  - Use User Embedding vector as the input
  - Use cosine similarity as affinity

- Validate User Grouping: Semantic Meaning
  - No ground truth.
  - Use **TF-IDF** to select representative words from each user group
    - W: a specific word, U: Tweets of a specific user group, D: Tweets of all users
    - TF(Term Frequency) - the occurrence of W in U
    - IDF - the inverse of the occurrence of W in D
    - TF*IDF
  - Need human annotation
    - Itfdb - it’s time for Dodgers Baseball
User Grouping

- Results
  - A user group of sightseeings in LA
  - Santa Monica and Venice are beachfront neighborhoods in LA
  - Beverly Hill and Hollywood are great tour places
  - flylaxairport - Lax Airport
  - This user group may be consist of travelers
User Grouping

• Results
  ◦ A user group of nightclubs and fashion brands
  ◦ elevennightclub, therasputin, playhouse, supperclub, and highrise are all nightclubs or nightclub events
  ◦ allaccesslifestyle, and hba are fashion brands
  ◦ People may wear fashion brands when they go to nightclubs

allaccesslifestyle
elevennightclub
tcny
therasputinweho
playhousehw
supperclub
hba
futboleros
highrisewho
pyrexvision
Hotspot Detection

- Using a variation of DBSCAN, but integrating **distance** and **similarity in words**
- The distance function is

\[
\Delta(i, j) = \alpha d(i, j) + \beta \frac{\vec{w}_i \cdot \vec{w}_j}{\|\vec{w}_i\|_2 \|\vec{w}_j\|_2}
\]

where \(\alpha, \beta\) are coefficients, and \(d\) is the distance function.

**Similarity in words:** Generate a vector for each word from the embedding, and calculate the average of the vector sum over the whole tweet.
Reasons of Choosing DBSCAN

1. It does not require one to specify the number of clusters in the data.
2. It can find arbitrarily shaped clusters. A hotspot in the map is not necessary a circle, since it can be a street, or multiple streets in one area.
3. One cluster can surround another one. (i.e. A chapel in the center of town)
Geographic Distance Function

1. 2-Norm between the coord pair

\[ d(i, j) = \|(x_i, y_i) - (x_j, y_j)\|_2 \]

2. Exponential function given distance

\[ d(i, j) = \exp(\lambda \|(x_i, y_i) - (x_j, y_j)\|_2) - 1 \]

3. Geographic distance between using formula (in meter)

\[
\text{hav} \left( \frac{d}{r} \right) = \text{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1) \cos(\varphi_2) \text{hav}(\lambda_2 - \lambda_1)
\]

\[
\text{hav}(\theta) = \sin^2 \left( \frac{\theta}{2} \right) = \frac{1 - \cos(\theta)}{2}
\]
Parameter tuning

\[ \Delta(i, j) = \alpha d(i, j) + \beta \frac{\vec{w}_i \cdot \vec{w}_j}{\|\vec{w}_i\|_2 \|\vec{w}_j\|_2} \]

DBSCAN parameters:
- \( \epsilon \), min_sample, \( \alpha \) and \( \beta \),

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Number of Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold Value</td>
<td>More than 20 (Any number between 0 to 0.1)</td>
</tr>
<tr>
<td>Distance function</td>
<td>3</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>More than 5</td>
</tr>
<tr>
<td>( \beta )</td>
<td>More than 5</td>
</tr>
</tbody>
</table>

After normalization, alpha is 3. Beta is 0.25. Epsilon = 0.045, min sample 5.
Results Checklist

1. Distinguish different landmarks in vicinity. (UCLA from Santa Monica.)
2. Distinguish local landmarks at the same place in different categories.
   a. Dodge Stadium from the shopping malls.
   b. Distinguish Staples Center from DTLA (downtown LA).
3. Distinguish small but concentrated cluster in nowhere. (Universal Studios,
   Griffith Observatory, Getty Center).

TF-IDF on the tweets in each cluster. Yields promising results.
Retrieve semantic meanings

1. Distinguish different landmarks in vicinity. (UCLA from Santa Monica.)
2. Distinguish local landmarks at the same place in different categories.
   a. Dodge Stadium from the shopping malls.
   b. Distinguish Staples Center from DTLA (downtown LA).
3. Distinguish small but concentrated cluster in nowhere. (Universal Studios, Griffith Observatory, Getty Center).
How to Find Mobility Patterns

- Frequent Triplet
  - Weekday, hour, cluster
  - Intuition: people tend to conduct certain event at a specific time
  - eg. go to club at night on the weekend
  - Chi-square independence test
    - Simply count the number of occurrence in each user group

- Trajectory
  - Transitions among hotspots
  - Generate candidates by imposing time and location constraints
    - eg. max time difference < 5 hr; location changes > 500m.
  - Use sequential mining techniques to output the most frequent trajectories.
Frequent Triplets Illustration

● Top Triplets of Group 0:
  ○ (Sat, 2, 11): hollywood, whiskeybluhw, joeybuicesound, supperclub, whiskeyblu
  ○ (Sat, 1, 11): hollywoodbowl, hollywood, tonight, whiskeybluhw, davestewart
  ○ (Sat, 21, 11): hollywood, night, event, find, game

● Top Triplets of Group 2:
  ○ (Tue, 22, 3): dodger, lakings, stadium, game, itfdb
  ○ (Fri, 22, 3): summerslam, wwe, wyatt, dodger, staplescenter
  ○ (Sat, 0, 3): dodger, night, lakings, chinatown, dtla

● Annotations of our data:
  ○ Hollywood: a lot's of bar, in our data, are located in west hollywood
  ○ Whiskeyblu: a bar with live music
  ○ Joeybuicesound: one event hosted in whiskeyblu in the period of our data
  ○ Hollywoodbowl: an amphitheatre with lots of shows at night
  ○ Itfdb: “It’s Time For Dodger Baseball!”
Trajectory Pattern Illustration

- **Group 0:**
  - 2->10: 2 means stadium. 10 means night club.
    - Eg: Dodger’s fans may go to nightclub to celebrate their winning.
  - 13->10/10->13: 10 means night club. 13 means hotels.
    - Eg: Travellers may go to characteristics local clubs.

- **Group 4:**
  - 1->10: 1 means santa monica beach. 10 means night club.
    - Eg: Travellers may go to beach first, and go to clubs at night.

- **Group 6:**
  - 2->59: 2 means stadium. 59 means neighborhoods.
    - Eg: Audience leave stadium and go back to their neighborhoods after games.
Future Work

- Employ more effective embedding technique
- Incorporate phrase mining in removing meaningless users and find semantic meanings for user groups
- Try a different distance function for Hotspot detection.
- Automatically generate semantic meanings for sequential trajectories.
Acknowledgement

We would like to thank Chao Zhang for providing and leading this project.

We would like to thank Keyang Zhang for providing help in technical challenges.

We would like to thank Professor Han and all TAs for providing this course.
Q&A

Any questions?
References

