Mobility Modeling and Prediction in Bike-Sharing Systems

Zidong Yang, Ji Hu, Yuanchao Shu, Peng Cheng, Jiming Chen and Thomas Moscibroda
Growing Dramatically Worldwide
Growing Dramatically Worldwide

> 500 bike-sharing systems
> 50 countries
> 1,000,000 shared bicycles
Background

What’s unique about bike-sharing?

<table>
<thead>
<tr>
<th>On-demand</th>
<th>Decentralized</th>
<th>Unattended</th>
<th>Concentrated</th>
</tr>
</thead>
</table>

Compared with other forms of shared-use mobility
1. Unlike conventional public transit (e.g., subways and buses) which follows a regular schedule and pre-determined routes, bike-sharing provides transportation on an on-demand basis with a decentralized structure.
2. Bike-sharing differs from classic ride-sharing (e.g., carpooling) and ride-sourcing (e.g., Uber and Lyft) in that bicycles are typically unattended. Also, during vacant hours, bicycles are concentrated at a group of stations.
Challenges
Challenges

Uneven distribution of bikes across stations

- Caused by uncontrolled, uneven usage demand
- Making check in or check out service unavailable at some stations
- Bike redistribution is non-trivial

Balancing Bike-Share Stations Has Become a Serious Scientific Endeavor

Some top mathematicians and computer scientists are devoting time to the problem.

ERIC JAFFE | @e_jaffe | Aug 27, 2014 | 43 Comments
Challenges

Uneven distribution of bikes across stations

- Caused by uncontrolled, uneven usage demand
- Making check in or check out service unavailable at some stations
- Bike redistribution is non-trivial
Challenges

Network modeling is the key and foundation
To understand how people rent and return bicycles
To understand how bicycles move among stations

Studies have been conducted
Extensive research on the nature of BSS, business models, how they have spread and adopted
Limited station clustering and coarse-grained rental volume forecasts

No fine-grained modeling and prediction
Main contributions

Spatio-temporal mobility model
To model the bike-sharing system as a dynamic network
To take into account the interactions among all stations

Traffic (check in/out) prediction mechanism
To jointly consider the spatio-temporal correlations and additional time factors and meteorology
On a per-station basis with sub-hour granularity

Evaluation with world’s largest public BSS
More than 2800 stations and over 103 million check in/out records
Best performance with an 85 percentile relative error of 0.6 for both check in and check out prediction
Design
Network modeling and flow prediction

Problem formulation

- Active objects (users) and Reactive objects (stations)
  - A shift instance (SI) = check out + movement + check in
- Coupled vs. Mutually independent
Network modeling and flow prediction

Design overview

• Modeling the mobility of undocked bicycles
  • Probabilistic model based on historical data to describe the bike movements

• Modeling the check out behaviors
  • Random forest theory to model and predict the check out behaviors

Time factors, meteorology

Station distribution, Bike availability

Check out
Riding, Check in
Network modeling

Theoretical mobility model

Aim to quantify bikes that will be checked in at station $i$ during target period $[t, t + \Delta t]$ in the future

\[
A_i = \sum_{j \in N} D_j \Gamma_{ji} P_t
\]

- $A_i$: The number of bikes checking in to station $i$
- $D_j$: The number of bikes checking out from station $j$
- $\Gamma_{ji}$: The transfer probability from station $j$ to station $i$
- $P_t$: The probability that the bike will check in to station $i$ within the target period
Network modeling

Theoretical mobility model

Temporal discretization

\[ n_{ji} = \sum_{k=1}^{\infty} D_j(t_k, \delta) Y_{ji}(t_k) \left( F_{ji}(t + \Delta - t_k) - F_{ji}(t - t_k) \right) \]

\( Y_{ji} \) and \( F_{ji} \) can be obtained based on historical SI data

Get the expression for \( A_i(t, \Delta t) = \sum_{j \in N} n_{ji} \)
Network modeling

Pruning

Temporal pruning
- 99.6% SIs are completed within 3 hours
- \( k \in [0, \infty) \Rightarrow k \in [0, 3] \)

Spatial pruning
- Top 200 stations contribute 96.6% of bikes on average
- \( N \leq 2800 \Rightarrow N \leq 200 \)

\( Y_{ji}(t) \) Discretization and Calculation
- Discretize \( Y_{ji}(t) \) into a piece-wise function
- Compute its value within each time slot (0.5 hour) based on historical check in/out data
Bicycle check out prediction

Feature extraction

Offline features

• Time factors (day of week, time of day, weekday, holiday)
• Meteorology (temperature, humidity, visibility, wind speed)

Online features

• Online check out number from the previous time window
Bicycle check out prediction

Random forest model

Deals with both categorical and numerical variables
Provides importance of features
Can be easily parallelized

Put it all together

- Check out volume
- Check out records from the past
- Predicted results in the future
- Mobility model
- Check in estimation
Evaluation
Dataset description

The BSS dataset
World’s largest public BSS in Hangzhou, China
Over 3300 stations and 84,000 shared bikes, 103,661,080 records

The meteorology dataset
Weather conditions of Hangzhou with 17,520 (i.e., 24*2*365) records

<table>
<thead>
<tr>
<th>user_id</th>
<th>rent_netid</th>
<th>tran_date</th>
<th>tran_time</th>
</tr>
</thead>
<tbody>
<tr>
<td>6114381</td>
<td>4051</td>
<td>20130101</td>
<td>000152</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>return_netid</th>
<th>return_date</th>
<th>return_time</th>
<th>bike_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>4015</td>
<td>20130101</td>
<td>001547</td>
<td>913672</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time (CST)</th>
<th>Temp (°F)</th>
<th>Dew Point (°F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:30 PM</td>
<td>100.4</td>
<td>69.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pressure (hPa)</th>
<th>Humidity (%)</th>
<th>Visibility (MPH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>29.65</td>
<td>37</td>
<td>6.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wind Dir</th>
<th>Wind Speed (MPH)</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSW</td>
<td>8.9</td>
<td>Partly Cloudy</td>
</tr>
</tbody>
</table>
Evaluation

Baseline approaches

- Historical Average (HA) [Gast et al. CIKM’15]
- Auto-Regressive and Moving Average (ARMA) [Vogel et al. CL’11]
- HP-MSI/P-TD [Li et al. SIGSPATIAL’15]

Evaluation methodology

- Check out prediction
- Check in estimation
Evaluation
Check out prediction

• Case study
  • Check out number over 24 hours
  • Summer > Winter
  • Different feature importance

<table>
<thead>
<tr>
<th>Day of week</th>
<th>Hour</th>
<th>Temperature</th>
<th>Humidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0288</td>
<td>0.1434</td>
<td>0.0846</td>
<td>0.0514</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Visibility</th>
<th>Wind speed</th>
<th>Holiday</th>
<th>Workday</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0332</td>
<td>0.0211</td>
<td>0.0030</td>
<td>0.0064</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Online check out number</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6282</td>
</tr>
</tbody>
</table>

(a) A rainy summer weekday.

(b) A sunny winter weekend.

Check out prediction at station 3648
Evaluation

Check out prediction

- **Overall performance**
  - First 20 days of each month to train, and predict the numbers in remaining days
  - Absolute error: numbers of bikes
  - Relative error: dividing the absolute error by the ground truth.
Evaluation

Check in estimation

• **Overall performance**
  
  • First 20 days of each month to train, and predict the numbers in remaining days
  
  • $\Delta = 30$ (i.e., each approach is required to estimate check in number in the following 30 minutes)
  
  • Best relative error from MM

---

**Overall performance**

(a) CDF of absolute error.

(b) CDF of relative error.
Insights

Variation Among Different Scenarios
Better prediction from workdays and stations in business area

User-centric Modeling and Prediction
Identify regular users and exploit their profiles
$n$ routes of which the check out/in stations fall into two small circles

<table>
<thead>
<tr>
<th>$n$</th>
<th>regular user(%)</th>
<th>$n$</th>
<th>regular user(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>12.55%</td>
<td>14</td>
<td>3.85%</td>
</tr>
<tr>
<td>8</td>
<td>9.19%</td>
<td>16</td>
<td>2.87%</td>
</tr>
<tr>
<td>10</td>
<td>6.65%</td>
<td>18</td>
<td>2.02%</td>
</tr>
<tr>
<td>12</td>
<td>5.02%</td>
<td>20</td>
<td>1.23%</td>
</tr>
</tbody>
</table>

Regular user percentage
Open research issues

From a mobile system point of view

Mobility Model Fusion with Multi-source Data
Exploit the inherent diversities from multi-source data (i.e., taxi, bus, subway)

Bike Rebalancing
Design an efficient and practical rebalancing algorithm

Service Optimizations
Station location optimization, service hour optimization, pricing strategy design etc.
Demo: Data Analysis and Visualization in Bike-Sharing System
Thank you

yuanchao.shu@microsoft.com