Real-Time Trip Information Service for a Large Taxi Fleet

Based on a paper by Rajesh Krishna Balan, Nguyen Xuan Khoa, and Lingxiao Jiang
Goal

- A system that uses historical taxi trip data to allow passengers to query the expected time and cost of a taxi trip that they plan to take.
- One taxi company in Singapore
Challenges

- Amount of data (tens of millions of records each month)
- Ability to answer queries in real time
- Accounting for various time-related factors (peak hours, highly variable taxi fare in Singapore).
- How much historical data to use
- How to filter out noise in data
Singapore taxi system

- 710 km² of area (37% larger than Warsaw)
- Densely populated - 5 million people (3 times more than in Warsaw)
- Taxis widely available and low priced
- ~25k taxicabs
- Ad-hoc pricing is not allowed
- Complicated charges
- Most pickups are street pickups
- Taxis are used for all activities
Data

- GPS in every taxi
- Start point, end point, distance, fare
- Intermediate points discarded
- 15k taxicabs, 35k taxi drivers
- 21 months
- 250 million trip records
- 3.6% trip records were anomalous (location errors, semantic errors)
10k random points from one day's data (0.3% one day's data)
Data

- Taxis were occupied 30% of the time
- Many trips with the same start and end place
Service requirements

- Accuracy (2 S$, 5 minutes)
- Real-time capability
- Low computational requirements (2 64G servers)
- Easy to deploy
Failed solution: Google Maps

- Network latencies and rate limits
- Problems with accuracy (about 40% errors)
- Local taxi trip prediction system (gothere.sg) had the same problems
Solution: trip history

- Basic features: start location, end location, start time
- Find *similar* trips and count their average
- PostgreSQL - took ~30 seconds to find trips that were similar enough
- Solution: splitting data into discrete partitions (time-space partitions)
Time windows partitioning

- Hourly Windows (HR)
- Day-of-Week Windows (DoW)
- Hourly DoW (DoW x HR)
- Peak period - splitting a day into 5 different periods with different charging (PEAK)
Static zoning

- Singapore fits into rectangle 25 km x 50 km
- Partition trips' start and end locations into squares (50 x 50, up to 5000 x 5000)
- Remove empty zones (unreachable or outside Singapore)
- Store average of trip details into hash map mapping selected type of time window and static zone to their prediction.
## Static zones

<table>
<thead>
<tr>
<th>Zone size (meters)</th>
<th>Total number</th>
<th>Number after compaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 x 50</td>
<td>565,586</td>
<td>162,730 (71%)</td>
</tr>
<tr>
<td>100 x 100</td>
<td>141,148</td>
<td>56,881 (60%)</td>
</tr>
<tr>
<td>150 x 150</td>
<td>62,559</td>
<td>31,834 (49%)</td>
</tr>
<tr>
<td>200 x 200</td>
<td>35,216</td>
<td>21,346 (39%)</td>
</tr>
<tr>
<td>250 x 250</td>
<td>22,374</td>
<td>15,285 (32%)</td>
</tr>
<tr>
<td>300 x 300</td>
<td>15,510</td>
<td>11,612 (25%)</td>
</tr>
<tr>
<td>350 x 350</td>
<td>11,502</td>
<td>9,197 (20%)</td>
</tr>
<tr>
<td>400 x 400</td>
<td>8,804</td>
<td>7,374 (16%)</td>
</tr>
<tr>
<td>450 x 450</td>
<td>6,930</td>
<td>6,017 (13%)</td>
</tr>
<tr>
<td>500 x 500</td>
<td>5,544</td>
<td>4,960 (11%)</td>
</tr>
</tbody>
</table>
Dynamic zoning

- Finding $k$ closest trips
- Start time is scaled according to average taxi speed
- Using kd-trees
- Still partitioning using time window
Evaluation methodology

- Dividing data into Set 1 (20 months) and Set 2 (1 month)
- History sets - incremental subsets of Set 1
- Set 2 used as query data for the system taught on different-sized history sets
Static zoning results - cost

Cost prediction better than expected

![Graph showing static zoning results - cost](image)

- LOC
- DOW
- PEAK
- HR
- DOW x HR

Average Prediction Error (cents)

Zone Size (metres)
Static zoning results - time

Average Prediction Error (secs)

Zone Size (metres)

LOC
PEAK
DOW
HR
DOW x HR
Static zone results - rate
Static zone results - rate

- 250m
- 200m
- 150m
- 100m
- 50m

Hit Rate (%) vs. Number of Months
Dynamic zoning results
Dynamic zoning over time
Performance comparison

Static zoning with DOW x HR and 200m zones

Dynamic zoning with k = 25
Accuracy analysis

- Indirect routes
- Traffic conditions
Anomalous trips

- Filter 1 - distance longer than 2 times straight line distance
- Filter 2 - average speed lower than 20 km/h or higher than 100 km/h
- Filter 1 - 9.5%
- Filter 1 + Filter 2 - 21%
Filter evaluation

![Graphs showing average prediction error over months for different filter settings.](image)
Traffic conditions

- Peak hours
- Special events in the city
- Weather, accidents
- Classifying trips according to weather. If the trip started in a zone where there has been enough rain AND ended in one.
- Only 0.6% classified as raining.
Weather impact on predictions
Summary of results

- Dynamic zoning with 6 months of data deemed best (with 0.9 S$ and 2.5 minute errors)
- Static zoning has too low hit rate
- Specific conditions as indirect routing and weather should be identified
Thank you!

Questions