Identifying Tourists from Public Transport Commuters

Mingqiang Xue ^
Huayu Wu ^
Wei Chen ^
Wee Siong Ng ^
Gin Howe Goh #

^ Institute for Infocomm Research
# Land Transport Authority of Singapore
Outline

• Introduction
• Background
• Our Approach
  – Station Ranking
  – Label Inference
• Experiments
• Case Study
• Related Work
• Conclusions
Introduction

• Tourism industry, a key economic driver for Singapore:
  – 15 million foreign visitors a year
  – 23 billion Singapore Dollar receipts in 2012

• Understanding tourists travelling behaviours is important:
  – Where do they go?
  – How they travel from one place to another?
  – Where do they stay?

• Useful to stake holders:
  – Government (tourism board, city planning, public transport): better planning, improve existing services
  – Private (travel agencies, taxis, hotels, restaurants, advertising etc): better or new business
Introduction

• A highly efficient transport system in Singapore
  – Buses, MRTs, LRTs
  – Payment mostly with commuter card (EZ-link)
  – Trajectories (partially) recorded

• Utilized by both locals, business travellers, and tourists in Singapore

• Who Are the Tourists Among the Commuters?
Introduction
Introduction

Main focus
Background – public transport

- The public transport system
  - MRT, similar to the subway in NYC
  - LRT, short distance neighborhood railway transport
  - Bus
Background – ticketing & Payment

Regular EZ-Link Card

Standard Ticket

Ticket by Cash

MRT

LRT

BUS
## Background – travel record

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Card_Number_E</td>
<td>Card ID for this ride</td>
</tr>
<tr>
<td>Transport_Mode</td>
<td>BUS, LRT, or MRT</td>
</tr>
<tr>
<td>Entry_Date</td>
<td>Date when ride started</td>
</tr>
<tr>
<td>Entry_Time</td>
<td>Time when ride started</td>
</tr>
<tr>
<td>Exit_Date</td>
<td>Date when ride ended</td>
</tr>
<tr>
<td>Exit_Time</td>
<td>Time when ride ended</td>
</tr>
<tr>
<td>Payment_Mode</td>
<td>Method of payment</td>
</tr>
<tr>
<td>Origin_Location_ID</td>
<td>Starting location of the ride</td>
</tr>
<tr>
<td>Destination_Location_ID</td>
<td>Ending location of the ride</td>
</tr>
</tbody>
</table>

The travel record Schema
Background

• Many tourists use standard tickets to travel around

• Tourists travelling patterns from standard tickets records
  – Problem: discontinued trajectories, no bus records, size could be small

• Our goal: identify tourists from regular EZ-link card users
Our Approach

- A Two staged process:
  - Stage 1: Initialization
    - Score each MRT/LRT station based on the attractiveness to tourists
  - Stage 2: Iterative Refinement
    - Update the scores for both MRT/LRT stations and tourists in a graph
    - Classify one as a tourist/non-tourist after the final iteration
Our Approach – Stage 1

• $t$ - a tourist commuter
• $m_i$ - an event that a commuter has visited station $i$

• We solve for each station:

$$\text{Score } s_{m_i} \sim \Pr(t|m_i)$$
Our Approach – Stage 1

• \( t \) - a tourist commuter
• \( m_i \) - an event that a commuter has visited station \( i \)

• We solve for each station:

\[
\text{Score } s_{mi} \sim \Pr(t|m_i) = \Pr(t) \cdot \frac{\Pr(m_i|t)}{\Pr(m_i)}
\]
Our Approach – Stage 1

• $t$ - a tourist commuter
• $m_i$ - an event that a commuter has visited station $i$

Score $s_{mi} \sim \Pr(t|m_i) = \Pr(t) \cdot \frac{\Pr(m_i|t)}{\Pr(m_i)}$
Our Approach – Stage 1

- $t$ - a tourist commuter
- $m_i$ - an event that a commuter has visited station $i$
- $n_i^s$ - number of trips with standard tickets at station $i$
- $n_i^r$ - number of trips with regular EZ-link card at station $i$
- $n_i^t$ - number of trips from tourists with standard tickets at station $i$

Score $s_{m_i} \sim \Pr(t|m_i) = \Pr(t) \cdot \frac{\Pr(m_i|t)}{\Pr(m_i)}$

The estimation of $\Pr(m_i|t)$:
- Idea: standard tickets records, but isolate the effects of locals
- $\hat{\Theta}$ is the probability that a local uses a standard ticket

$$\hat{\Pr}(m_i|t) = \frac{n_i^t}{\sum_i n_i^t} \text{ where } n_i^t = n_i^s - n_i^r \cdot \hat{\Theta}$$
Our Approach - Stage 1

- The estimation of $\hat{\theta}$:

<table>
<thead>
<tr>
<th>Name</th>
<th>$n_i^s$</th>
<th>$n_i^r$</th>
<th>$\frac{n_i^s}{n_i^r}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marymount</td>
<td>6218</td>
<td>629435</td>
<td>0.009879</td>
</tr>
<tr>
<td>Yio Chu Kang</td>
<td>20361</td>
<td>2067636</td>
<td>0.009847</td>
</tr>
<tr>
<td>Cove</td>
<td>1817</td>
<td>189873</td>
<td>0.009570</td>
</tr>
<tr>
<td>Buangkok</td>
<td>7454</td>
<td>787463</td>
<td>0.009466</td>
</tr>
<tr>
<td>Layar</td>
<td>345</td>
<td>37211</td>
<td>0.00927</td>
</tr>
<tr>
<td>Oasis</td>
<td>489</td>
<td>53696</td>
<td>0.009107</td>
</tr>
<tr>
<td>Labrador Park</td>
<td>2473</td>
<td>292858</td>
<td>0.008444</td>
</tr>
<tr>
<td>Tongkang</td>
<td>1295</td>
<td>158299</td>
<td>0.008181</td>
</tr>
<tr>
<td>Compassvale</td>
<td>2705</td>
<td>358175</td>
<td>0.007552</td>
</tr>
<tr>
<td>Dover</td>
<td>8963</td>
<td>1247247</td>
<td>0.007186</td>
</tr>
</tbody>
</table>
Dover surroundings: - An isolated educational institution
- No closeby residences
Our Approach – Stage 1

- $t$ - a tourist commuter
- $m_i$ - an event that a commuter has visited station $i$
- $n_i^s$ number of trips with standard tickets at station $i$
- $n_i^r$ number of trips with regular EZ-link card at station $i$
- $n_i^t$ number of trips from tourists with standard tickets at station $i$

Score $s_{m_i} \sim \Pr(t|m_i) = \Pr(t) \times \frac{\Pr(m_i|t)}{\Pr(m_i)}$

The estimation of $\Pr(m_i)$:

$$\hat{\Pr}(m_i) = \frac{n_i^s + n_i^r}{\sum_i n_i^s + n_i^r}$$
Our Approach – Stage 1

- \( t \) - a tourist commuter
- \( m_i \) - an event that a commuter has visited station \( i \)
- \( n_i^s \) number of trips with standard tickets at station \( i \)
- \( n_i^r \) number of trips with regular EZ-link card at station \( i \)
- \( n_i^t \) number of trips from tourists with standard tickets at station \( i \)

Score \( s_{m_i} = \Pr(t) \cdot \frac{\Pr(m_i|t)}{\Pr(m_i)} \)

where \( \hat{\Pr}(t) = \frac{\sum_i 2n_i^t}{\sum_i n_i^s + n_i^r} \)
## Our Approach – Stage 1

<table>
<thead>
<tr>
<th>Name</th>
<th>$s_{m_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Changi Airport</td>
<td>0.213668</td>
</tr>
<tr>
<td>Marina Bay</td>
<td>0.145012</td>
</tr>
<tr>
<td>Clarke Quay</td>
<td>0.144702</td>
</tr>
<tr>
<td>Bayfront</td>
<td>0.128008</td>
</tr>
<tr>
<td>Little India</td>
<td>0.118879</td>
</tr>
<tr>
<td>Chinatown</td>
<td>0.113837</td>
</tr>
<tr>
<td>HarbourFront</td>
<td>0.106443</td>
</tr>
<tr>
<td>Bras Basah</td>
<td>0.104787</td>
</tr>
<tr>
<td>Esplanade</td>
<td>0.099637</td>
</tr>
<tr>
<td>Orchard</td>
<td>0.098623</td>
</tr>
<tr>
<td>Lavender</td>
<td>0.093104</td>
</tr>
<tr>
<td>Farrer Park</td>
<td>0.081844</td>
</tr>
<tr>
<td>Promenade</td>
<td>0.079080</td>
</tr>
<tr>
<td>Bugis</td>
<td>0.070973</td>
</tr>
<tr>
<td>City Hall</td>
<td>0.064815</td>
</tr>
</tbody>
</table>

Top Ranked stations based on attractiveness
Our Approach – Stage 2

A toy Station-Commuter Relationship graph
Our Approach – Stage 2

• While # of iterations < predefined threshold (e.g. 150):

  – Update the class distribution of each commuter based on its current class distribution and the class distributions of stations that they visited

  – Update the class distribution of each station based on its current distribution and the class distributions of commuters who visit them
Our Approach – Stage 2

[Diagram showing interconnected nodes and edges with labels: t1, t2, t3, t4, I1, I2, I3, I4, Harbour Front, Changi Airport, City Hall, Oasis, Dover. Numbers and coordinates are also indicated for each node.]
Our Approach – Stage 2
Our Approach – Stage 2

- Updating functions:

\[ \phi_{l_i}^k \leftarrow \alpha \cdot \phi_{l_i}^{k-1} + (1 - \alpha) \cdot \frac{\sum_{m \in N(l_i)} w_{l_i m} \cdot \phi_{m}^k}{\sum_{m \in N(l_i)} w_{l_i m}} \]

\[ \phi_{t_i}^k \leftarrow \beta \cdot \phi_{t_i}^{k-1} + (1 - \beta) \cdot \frac{\sum_{m \in N(t_i)} w_{t_i m} \cdot \phi_{m}^k}{\sum_{m \in N(t_i)} w_{t_i m}} \]

\[ \phi_{m_i}^k \leftarrow \gamma \cdot \phi_{m_i}^{k-1} + (1 - \gamma) \cdot \frac{\sum_{u \in N(m_i)} w_{um_i} \cdot \phi_{m_i}^k}{\sum_{u \in N(m_i)} w_{um_i}} \]

Update for commuters

Update for stations
Our Approach – Stage 2

• Final class assignment:

\[ \hat{C} = \arg\max_c \frac{P(t_i | c)}{P(t_i)} = \arg\max_c \frac{P(c | t_i)}{P(c)} \]

For \( c \in \{\text{Tourist, Non−Tourist}\} \)
Experiments

• One-month EZ-link records from LTA

• Preprocessing:
  – Exclude commuters with less than 6 records

• Data description:
  – 1.7 million commuters
  – 49.5 million records
  – Training set: 1000 tourists and 250,000 locals

• Competitors:
  – FTF (Fast Transversal Filter): a state-of-the-art iterative inference algorithm
  – SVM

• Evaluation metric:
  – F1 score: \( F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Recall} + \text{Precision}} \)
### Experiments

#### Comparison Results

<table>
<thead>
<tr>
<th>p%</th>
<th>SVM Macro F1</th>
<th>SVM Micro F1</th>
<th>FTF Macro F1</th>
<th>FTF Micro F1</th>
<th>I² Macro F1</th>
<th>I² Micro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>0.57984</td>
<td>0.8415</td>
<td>0.6109</td>
<td>0.8419</td>
<td>0.6267</td>
<td>0.8504</td>
</tr>
<tr>
<td>10%</td>
<td>0.5917</td>
<td>0.8420</td>
<td>0.6263</td>
<td>0.8464</td>
<td>0.6572</td>
<td>0.8538</td>
</tr>
<tr>
<td>15%</td>
<td>0.6144</td>
<td>0.8411</td>
<td>0.6441</td>
<td>0.8433</td>
<td>0.6677</td>
<td>0.8560</td>
</tr>
<tr>
<td>20%</td>
<td>0.6199</td>
<td>0.8480</td>
<td>0.6758</td>
<td>0.8504</td>
<td>0.6962</td>
<td>0.8575</td>
</tr>
<tr>
<td>25%</td>
<td>0.6286</td>
<td>0.8402</td>
<td>0.6956</td>
<td>0.8459</td>
<td>0.7154</td>
<td>0.8549</td>
</tr>
</tbody>
</table>
Case Study

Places visited by tourists by popularity
Case Study

Where do tourists go from the airport?
Where do tourists go from Bugis?
Why do tourists visit Ang Mo Kio?
Related Work

• Mining public transport data
  – Improve public transport in a city
  – Behaviors of populations (what’s the popular shopping places)
  – Behaviors of individuals (what’s one’s home, work place)

• Mining tourists data
  – Travelling patterns of tourists (e.g based on Geo-tagged images)
Conclusions

• Extract tourists records from public transport data
  – Meaningful to stakeholders, both private and government

• Proposed an algorithm based on:
  – Station scoring and iterative score refinement

• Verified findings with experiments

• Hope to attract interest to solve similar problems in other cities, e.g. Hong Kong, NYC, London etc.
Thank you