Understanding semantic similarity among subway stations using smart card data

Reporters: Zhuang Dingyi, Siyu Hao
Supervisors: Prof. Lee Der-Horng, Prof. Jiangang Jin
Department of Civil and Environmental Engineering
National University of Singapore
Contents

1. Introduction
2. Research ideas
3. Research results
4. Discussion and analysis
5. Conclusion
1. INTRODUCTION

Motivation
Literature review
Contribution
Motivation

Country and Capital Vectors Projected by PCA

Words → Embedding → Word vectors → Clustering → Semantic similarity → Stations?

Sequential features
Literature review

Previous station similarity analysis is based on shallow mobility features, such as aggregated passenger flow

Mohamed, K.et. al. in Clustering smart card data for urban mobility analysis.

Some transferred semantic models into urban computing, but regarding stations as documents and lack of further comprehensive analysis


Semantic models are now widely applied in fields outside Natural Language Processing

Contribution

Concept

Stations are like Chinese characters or compound words

Meaning in sentence (Mobility pattern)

Words (Stations)

Literal meaning, e.g. superman=super+man (Inherent features like POI)

Case studies

Analysis on similarity between MRT stations of Singapore in a planning perspective:

- 9 POI categories
- 5 case studies
- Planning suggestions
2. RESEARCH IDEAS

Expected results
How to transfer
Expected results

words → Embedding → WordVectors → Clustering → Synonym/Antonym

Stations → Embedding → Semantic Vectors → Clustering → Similarity/Dissimilarity

Understanding & Planning → Hidden Correlation → Interpretation
How to transfer

Compound Words

Meaning in context

Literal meaning, e.g. superman = super + man

Mobility pattern (Inflow & outflow)

Stations

Service feature (POI)

Proposed steps:

Inflow & outflow

POI distribution

Autoencoder Embedding

TF-IDF

SVD

Service semantics

Mobility semantics

Hidden relationship

Urban Planning

Service category
3. RESEARCH RESULTS

Dataset
Stacked autoencoder
Mobility semantics
Service semantics
Case studies

3. RESEARCH RESULTS
Dataset

Provided by Land Transport Authority (LTA), Singapore. Multi-model data (Bus&MRT), we only considered MRT.

Table for dataset description

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covered days</td>
<td>2012/3/19-2012/3/25 (Normal week)</td>
</tr>
<tr>
<td>Covered Stops</td>
<td>4702 (122 for MRT stations)</td>
</tr>
<tr>
<td>Average records number each day</td>
<td>&gt;5,000,000</td>
</tr>
<tr>
<td>Data volume</td>
<td>4.1 GB</td>
</tr>
<tr>
<td>Average multi-model riding distance</td>
<td>7 km</td>
</tr>
<tr>
<td>Average multi-model riding time</td>
<td>20 min</td>
</tr>
<tr>
<td>Multi-model transferring percentage</td>
<td>30%</td>
</tr>
<tr>
<td>Average MRT riding distance</td>
<td>12 km</td>
</tr>
<tr>
<td>Average MRT riding time</td>
<td>27 min</td>
</tr>
<tr>
<td>MRT transferring percentage</td>
<td>23%</td>
</tr>
</tbody>
</table>
To better understand temporal influence, inspired by Mohamed, K.et al. in Clustering smart card data for urban mobility analysis. We choose 1-3 hours as our time interval (LOW VARIABILITY).
We divide the time into 7 time intervals:

- 5-7 pre-morning peak
- 7-10 morning peak
- 10-16 morning off-peak
- 16-17 pre-evening peak
- 17-19 evening peak
- 19-22 late evening peak
- 22-24 evening off-peak

Mobility vectors of same time intervals: m*n dimension, where

m: 122*7 = 854 (122 stations * 7 days)

n: 122+122+7 = 251 (inflow&outflow from&to all stations + one-hot code for day)
Dataset

POI dataset is powered by Google Maps, contains:

• 22 categories
  • 'atm','bank','bus_station', 'transit_station', 'place_of_worship'
  • 'supermarket', 'shopping_mall', 'education' , 'parking', 'park',
  • 'political', 'storage', 'intsec','lodging','hospital','car_rental',
  • 'car_dealer','car_repair','bar','cafe','local-government_office','bicycle_store'

• 10 MRT lines
  • 'NS','EW','NE','CC','CE','BP','CG','PE','SW','SE'
Stacked autoencoder

Reduce the dimension of flow vectors from 251 into 16. Train 7 models for 7 time intervals respectively. Train data use Min-Max normalization.

\[
\text{Loss} = \frac{(output - input)^2}{sampleSize}
\]
Stacked autoencoder

Platforms and training parameters:
• Epoch: 200000, batch size: 128, optimizer: adaGradient (LR: 0.01)
• 8 E5 cores, 16GB RAM, 1060 3GB, take 7 hours to train one model

<table>
<thead>
<tr>
<th>Time interval</th>
<th>R-squared value</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre-morning peak</td>
<td>0.881</td>
</tr>
<tr>
<td>morning peak</td>
<td>0.951</td>
</tr>
<tr>
<td>morning off-peak</td>
<td>0.959</td>
</tr>
<tr>
<td>pre-evening peak</td>
<td>0.882</td>
</tr>
<tr>
<td>evening peak</td>
<td>0.948</td>
</tr>
<tr>
<td>late evening peak</td>
<td>0.947</td>
</tr>
<tr>
<td>evening off-peak</td>
<td>0.865</td>
</tr>
<tr>
<td>Mean</td>
<td>0.919</td>
</tr>
</tbody>
</table>
Mobility semantics

Country and Capital Vectors Projected by PCA

"Capital" & "Country" ~ Mobility semantics in Different time intervals

Variability within same day

Variability within same tint

Mobility semantic vector decomposition
Mobility semantics

For semantic vectors Beijing-China≈Tokyo-Japan

Cosine similarity

Stn1_Monday_MorningPeak-Stn1_Monday_EveningPeak ≈ Stn2_Friday_MorningPeak-Stn2_Friday_EveningPeak

Elements similarity of each two time interval group’s subtraction vector
Service semantics

Term Frequency–Inverse Document Frequency (TF-IDF)

$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

SVD to find semantics (refer to literatures)
Categories are hard to classify, use topic modeling to help us better find the division.
## Service semantics

<table>
<thead>
<tr>
<th>Words</th>
<th>Clusters</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>atm</td>
<td>Cluster 0</td>
<td>1.3696</td>
<td>1.8781</td>
</tr>
<tr>
<td>bank</td>
<td>Cluster 0</td>
<td>1.1391</td>
<td>1.1029</td>
</tr>
<tr>
<td>bus station</td>
<td>Cluster 0</td>
<td>1.3124</td>
<td>2.2211</td>
</tr>
<tr>
<td>transit station</td>
<td>Cluster 0</td>
<td>1.3558</td>
<td>2.453</td>
</tr>
<tr>
<td>place of worship</td>
<td>Cluster 0</td>
<td>1.3757</td>
<td>1.2871</td>
</tr>
<tr>
<td>supermarket</td>
<td>Cluster 0</td>
<td>1.0753</td>
<td>1.3188</td>
</tr>
<tr>
<td>shopping mall</td>
<td>Cluster 0</td>
<td>1.0883</td>
<td>1.0688</td>
</tr>
<tr>
<td>education</td>
<td>Cluster 0</td>
<td>1.5323</td>
<td>2.2542</td>
</tr>
<tr>
<td>parking</td>
<td>Cluster 0</td>
<td>1.0407</td>
<td>1.1003</td>
</tr>
<tr>
<td>park</td>
<td>Cluster 0</td>
<td>0.9873</td>
<td>1.2058</td>
</tr>
<tr>
<td>political</td>
<td>Cluster 0</td>
<td>1.0004</td>
<td>1.0501</td>
</tr>
</tbody>
</table>

The results might change occasionally since samples are small.
Service semantics

POI clusters distribution

c0 developed residential area
c1 under-construction
c3 convenient traffic hub/old residential area?
c6 developed commercial regions
c7 potential planning regions
c8 entertainment
c9 scientific&educational
c10 emerging residential area
c11 emerging commercial regions
Case studies

1. Different lines, same POI semantics, same flow semantics
Case studies

1. Different lines, same POI semantics, same flow semantics (dL_sP_sF).

Discovered stations are usually LRT or other remote stations, because they same interaction station. Like Farmway and Woodleigh (C10 emerging residential area), might both share similar flow patterns from Sengkang.
Case studies

2. Same line, same POI semantics, same flow semantics (sL_sP_sF)

Flow similarity for stations in the same line with same POI category

This benefits best to advertisers. Discovered stations are usually the adjacent stations in the same line, such as Somerset and Orchard or Pioneer and Bonn Lay.
Remote stations in the same line, like Pasir Ris and Dover. While stations in residential region like Jurong East and Buona Vista are intersections to connect flow demand from different places.
Circle line and LRT lines are the most typical since they serve only particular regions. POI are quite different in the opposite sides but customer flow remains similar.
This result satisfies our knowledge, since distant stations in the same line serve different needs and located in various circumstances.
4. DISCUSSION AND ANALYSIS

Commercial interests
Urban planning
Further work
Commercial interests

Advertisement. Advertisers can focus on stations with the same POI & flow feature and avoid targeting stations with different POI & flow feature. In general, advertising among adjacent stations in the same line.

Site selection. For small and medium-size enterprises targeting at regular or similar customers, like cheap clothing stores, snack bars or barber shops, can refer to stations with the same flow features to develop core customers.
Urban planning

**Infrastructure.** Lanes, bus stops, etc. can be constructed according to same flow features or same POI, like Tampines and Jurong East (highest overlapping in sL_sP_sF).

**Traffic monitoring.** Crowd with similar boarding or alighting patterns can provide insight to understand customers mobility for emergent evacuation, especially for circle line.

**Land use.** Flow and POI relationship, no matter similar or not, could provide comprehension of urban land use. Low utilized stations, like Ten Mile Junction, Farmway and Woodleigh can be abolished for better land use.
Further work

**POI category division.** Our service semantics only gives a roughly divided POI categories, but sophisticated division might be further analyzed.

**Bus stops consideration.** We only focused on MRT stations, which, however, is only part of the public transportation system.
5. CONCLUSION

Highlights
Timeline
Highlights

• Transplant semantic models on urban mobility discovery
• Proposed a new comprehension of semantic model
• Discovering specific relationship between MRT stations
• Give solid urban planning analysis and suggestions
Timeline

- **18th - 24th July**: Feature engineering
- **10th - 17th July**: Literature reviewing
- **25th July - 5th Aug**: Model training
- **6th - 12th Aug**: POI processing Interpretation
- **13th - 22nd Aug**: Finish all the work
- **23rd - 27th Aug**: PPT summarization
THANK YOU