Data-Driven Customer Segmentation and Personalized Information Provision in Public Transit

Abhishek Arunasis Basu[1], Jinhua Zhao[1], Haris N. Koutsopoulos[2], Rabi G. Mishalani[3]

Abstract

The goal of this research is to develop a framework that could help a transit agency to better understand its users and their behaviors through the use of smart card data. The framework developed in this study includes a data-driven segmentation technique that can help identify customer groups based on the spatial and temporal aspects of their travel pattern. Data from Hong Kong's MTR system were used to demonstrate the effectiveness of the developed segmentation methodology and its value for practical applications. For example, this study demonstrated the feasibility of a personalized information provision system, which would be especially useful in the event of a service disruption.

Keywords

Data-driven segmentation, Unsupervised Learning, Personalized Information provision, Spatial and Temporal features, Smart Card data

Introduction

The public transportation industry has witnessed a data boom in the past decade through data collection from automated systems, with transit smart cards playing a key source of rich disaggregate data on customers' travel patterns. In this study, we utilized the disaggregate data obtained from smart cards for extraction of dominant travel patterns as recognized through grouping together users based on similar travel frequency, spatial, and temporal aspects of travel behavior. These homogeneous groups or customer segments utilize the transit network in similar ways. Understanding the characteristic traits of these segments can help the transit agency to better cater to their specific needs, like providing personalized information.

Data

Hong Kong's smart card system, Octopus, is the main source of data used in this study. A subset of these data pertaining to customer transactions on the MTR network was made available for the years 2015 and 2016. Only cards marked as 'Adult' were used in this study. This data covered travel on MTR's 9 heavy rail lines – the Island Line, Kwun Tong Line, Tsuen Wan Line, Tseung Kwan O Line, Tung Chung Line, Airport Express Line, East Rail Line, Ma On Shan Line, and West Rail Line.

Methodology and Analysis

The framework developed in this study consisted of a tiered mechanism for data-driven segmentation, and was defined separately for weekdays and weekends. All the analyses carried out pertain to the former. Results for the latter could be arrived at by applying the same framework.

The first tier is based on two features, defined as Spatial Probability that measured the customer's tendency to use a given station for their trips either as the origin or as the destination, and Temporal Probability that measures the customer's tendency to travel at a particular hour of the day. The features the second tier is based on relate to travel frequency, and relative spatial, and temporal characteristics. A total of 26 features are defined in this tier. Some features include the number of unique stations a customer visits, the number of days they are active on the network, the number of unique origin-destination pairs they travel on, the number of trips that begin or end at their inferred home station, to name a few.

[1]Massahusetts Institute of Technology
[2]Northeastern University
[3]The Ohio State University

Based on the above features, each customer is assigned one spatial short-term segment and one temporal short-term segment, with short-term signifying that these characteristics of the customers could potentially change in the short run, based on the features defined in the first tier. Based on the second tier, customers that have sufficient Octopus based data associated with them (in terms of length of time and frequency they've traveled on the network) would be assigned a long-term segment. The main technique used for identify segments in this study involved clustering through k-means++ with Davies Bouldin index as the criterion for selecting the number of clusters. Further, for long term segments, the stability of these segments is analyzed to ensure that the segments identified by using data from one time period are generalizable across time.

Results and Conclusions

In the first tier, based on the criterion described above, 15 short-term temporal segments were identified. These segments consisted of customers who tend to travel at specific periods of the day, which may be indicative of the type of activity they may engage in. For example, considering segment T1, it was inferred that it may be composed of office-goers who possess little flexibility in their start times, which is around 9 AM, but the time of return trip varied, sometimes it occurred at around 5 PM and at other times around 8 PM.

Similarly, a total of 24 short-term spatial segments were obtained. In this case, each segment consisted of customers who travel to and from a specific set of stations in the MTR network. For example, it was found that customers in segment S4 access stations that lie on either side of the harbor, on the Tsuen Wan Line and the Island Line respectively, with the stations Tsim Sha Tsui, Causeway Bay, East Tsim Sha Tsui, Wan Chai, and Jordan, being frequented the most.

In the second tier, a total of eight segments were obtained. Each segment, in this case, represents a group that exhibits homogeneous travel behavior in terms of frequency, relative spatial, and temporal patterns For example, it was found that Segment H consisted of users who rarely traveled more than once on a given day, and traveled over relatively short distances. Moreover, most of these trips began or terminated at a particular station on the network, which could be the location of their home or work (or other activity). While the first trip would occur around 9 AM in the morning, the time of last trip (if there was a second trip) would be highly variable. Therefore, these users would usually depend upon other modes of transport for their return trips. Further, stability analysis across data from years 2015 and 2016 showed that the segments obtained were indeed stable, and thus the results were generalizable through time.

The knowledge of segments obtained from the above methodology could be used to identify potential customers who may benefit from certain types of information. In this study, information provision during service disruptions was analyzed. It was found that by applying the developed segmentation framework, the transit agency could provide targeted information to customers who would be affected due to a disruption on a priority basis. Retrospective analysis showed that for a particular disruption event, the application of the framework allowed reaching 81% of affected customers and effectively reduce spam by 75%, with parameters that can be tuned based on the criticality of the information to be provided.

In conclusion, the segmentation scheme created in this study was found to be useful, interpretable and generalizable across time. It was shown to be effective for information provision, and could be extended to other applications such as prediction of customer attrition and for analysis of before-and-after effects of network expansion.