Inferring trip destinations in transit smart card data using a probabilistic topic model

Zhanhong Cheng[1], Martin Trépanier[2,3], Lijun Sun[1,3*]

[1]Department of Civil Engineering and Applied Mechanics, McGill University

[2]Polytechnique Montréal

[3]Interuniversity Research Center on Enterprise Networks, Logistics and Transportation (CIRRELT)

Abstract

When estimating transit trip destination by smart card data, traditional algorithms seek information from individuals' consecutive or historical trips. These algorithms cannot handle isolated unseen trips. This paper proposes a destination estimation algorithm by Latent Dirichlet Allocation (LDA) model, which can borrow information from travellers with similar travel pattern to infer the probabilities of possible destinations. A preliminary test in Guangzhou Metro system shows around 60% accuracy for the proposed model.

Keywords: smart card data, destination estimation, topic model

Introduction

Origin and Destination (OD) Matrix is an essential input for public transit planning and operation. Most transit agencies have been relying on travel surveys to collect representative OD information; however, conducting such a survey with reasonable scale is not only expensive but also time-consuming. With recent advances in ICT, researchers and practitioners have started taking advantage of the transit operation data and smart card data for better planning and operation (Pelletier et al., 2011).

In general, smart card systems are designed for the purpose of automatic fare collection. When the system has both tap-in and tap-out controls (e.g., using a distance-based transit fare scheme), the full itinerary (boarding time/station and alighting time/station) of each trip will be registered. However, most existing systems across the world adopt a single fare scheme with only tap-in control, and thus alighting information (time/station) is essentially unknown. To obtain accurate OD matrix from these systems, it is critical to impute the exact alighting station/stop of each transit trip.

In the literature, there have been several studies addressing this problem. Barry et al. (2002) proposed an algorithm based on two assumptions on passengers: (1) the destination station of one's previous trip has a high chance to be (or close to) the origin station of the next trip; (2) the last destination station of a day is often the first origin station of the same day. Base on similar assumptions, many empirical studies have been conducted in different cities (Trépanier et al., 2007; Munizaga and Palma, 2012; Nunes et al., 2016). Depends on the data, current algorithms can obtain around 60% to 80% destinations among all trips; the rest trips are mostly non-routine or discontinuous. To better infer the missing destinations, He and Trépanier (2015) used the kernel density probability of passengers' spatial and temporal feature and get an additional 10% accuracy. The idea is to fully utilize the information of a passenger's historical trips and infer the destination of trips based on previous trips with similar origin and departure time.

Despite the minor differences in terms of data sets, most of state-of-the-practice destination estimation algorithms are still based on the same assumptions as in Barry et al. (2002). Essentially, there are two major limitations in these algorithms: on the one hand, these algorithms are individual based and the application requires a specific model for each individual; on the other hand, these models require large amount of training data for each individual and the performance is limited when we have insufficient historical observations. To address these two issues, we propose

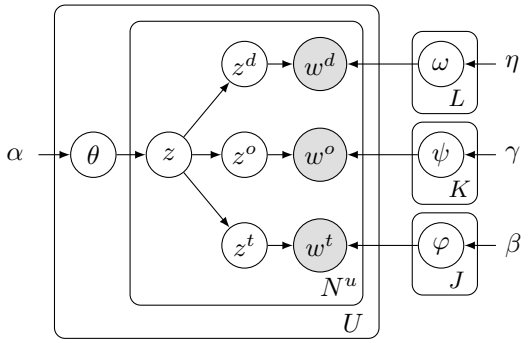

Figure 1: Plate diagram for the model

a collaborative filtering approach by sharing information from passengers with similar travel patterns. In detail, we adapt the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003) for individual transit smart card data by considering each passenger a corpus and each trip a word. LDA is a probabilistic generative model that firstly used in natural language processing to analyze the topics of articles (topic model). Recent years, this model has also been used to model the behavioral pattern (Hasan and Ukkusuri, 2014; Fan et al., 2016). Similar to the two-dimensional LDA developed by Sun et al. (2019) which was used to detect anomaly in travel behavior, we extend this model to a three-dimensional LDA to capture the trip character of transit system. The advantage of this method is that it avoids introducing a large vocabulary set and keeps the spatiotemporal relationship with in the latent space. Based on this model, we can estimate the posterior probability of destination stations of a trip conditional on the origin station and the boarding time.

We use the metro smart smart card collected from Guangzhou, China over three months for numerical experiments. The metro system in Guangzhou has both tapping-in and tapping-out controls and thus full itinerary (both origin and destination) information is available. In the experiment, we choose 1,000 passengers with a total number of trips between 50 and 200. We randomly select 70% of the data to fit the model and wipe out the destination for the rest 30% data to test the model. Our results show that the average prediction accuracy is about 60%, and the accuracy can be even higher for regular commuters with routine travel patterns.

Methodology

Topic model assumes there are a number of latent topics among documents, each topic has different word distribution and each document has different topic distribution. The methodology here is analogous to topic model. As each trip comprises time, origin, and destination three properties, the difficulty of adopting LDA in destination estimation lies in how to properly define the "word" and the "topic".

We use w^t, w^o, and w^d to represent the time, origin, and destination of a trip. Denote $\boldsymbol{w}_u = \{(w_i^t, w_i^o, w_i^d) : i = 1, \ldots, N_u; w_i^t \in \{1, \ldots, T\}; w_i^o, w_i^d \in \{1, \ldots, S\}\}$ to be the trip set of passenger u. Where N_u is the total number of known trips of passenger u, T and S are the size of temporal and spatial dimension respectively. Similar to the topic model, the trip set $\boldsymbol{w}_u$ is a document, each tuple (w_i^t, w_i^o, w_i^d) is a word. In order to let spatial and temporal dimension to have different number of topics, as well as maintain the relation between spatial and temporal topic. We organize the latent topic in the manner of a three-dimensional tensor $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{J \times K \times L}$, where the J, K, and L are the the number of latent topic in time, origin, and destination respectively. The element $z_{j,k,l}$ of tensor $\boldsymbol{\mathcal{Z}}$ corresponds to the j^{th} temporal topic z_j^t, the k^{th} origin topic z_k^o, and l^{th} destination topic z_l^d. And there are three set of topic-word distribution for time, origin, and destination dimension. A plate diagram for our model is shown in Figure 1. The generative process shown in Figure 1 is listed as follows:

- Draw topic distribution for passenger $\theta_u \sim \text{Dirichlet}(\boldsymbol{\alpha})$
- Draw time distribution for each temporal topic $\varphi \sim \text{Dirichlet}(\boldsymbol{\beta})$
- Draw spatial distribution for each boarding location topic $\psi \sim \text{Dirichlet}(\boldsymbol{\gamma})$
- Draw spatial distribution for each alighting location topic $\omega \sim \text{Dirichlet}(\boldsymbol{\eta})$

- For each passenger u, for each trip record:
  - Draw a topic $z \sim \text{Multinomial}(\theta_u)$
  - Let $z^d = \lceil z/(J \times K) \rceil$
  - Let $z^o = \lceil (z - (z^d - 1) \times (J \times K))/J \rceil$
  - Let $z^t = z - (z^d - 1) \times (J \times K) - (z^o - 1) \times J$
  - Draw $w^t \sim \text{Multinomial}(\varphi_{z^t})$
  - Draw $w^o \sim \text{Multinomial}(\psi_{z^o})$
  - Draw $w^d \sim \text{Multinomial}(\omega_{z^d})$

The probability of a user take a trip (w^t, w^o, w^d) is

$$P(w^t, w^o, w^d) = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} P(w^t|z_j^t) P(w^o|z_k^o) P(w^d|z_l^d) P(z_j^t, z_k^o, z_l^d). \tag{1}$$

A group of (z^t, z^o, z^d) corresponds to one latent pattern that jointly depicts the spatio-temporal pattern of a passenger. The same with Griffiths and Steyvers (2004), a collapsed Gibbs sampling algorithm was applied to infer the three-dimensional LDA model. The conditional topic distribution is as follows.

$$\begin{aligned} P(z_i^t = j, z_i^o = k, z_i^d = l | w_i^t = t, w_i^o = o, w_i^d = d, \boldsymbol{z}_{-i}^t, \boldsymbol{z}_{-i}^o, \boldsymbol{z}_{-i}^d, \boldsymbol{w}_{-i}^t, \boldsymbol{w}_{-i}^o, \boldsymbol{w}_{-i}^d) &\propto \varphi_{t,j} \cdot \psi_{o,k} \cdot \omega_{d,l} \cdot \theta_{u,j,k,l} \\ &= \frac{N_{z^t=j}^{w^t=t} + \beta}{N_{z^t=j} + T\beta} \times \frac{N_{z^o=k}^{w^o=o} + \gamma}{N_{z^t=k} + S\gamma} \times \frac{N_{z^d=l}^{w^d=d} + \eta}{N_{z^t=l} + S\eta} \times \frac{N_{z^t=j, z^o=k, z^d=l}^{u} + \alpha}{N^u + JKL\alpha}. \end{aligned} \tag{2}$$

Where $N_{(\cdot)}^{(\cdot)}$ denotes the number of trips when satisfying the condition listed in the subscript and the superscript. Note that the current trip i is excluded when counting N.

Having all the parameters in this model, we can infer the missing destination. The probability for passenger u alighting at a location d given the trip time t and boarding location o is given by Eq. (3) and can be calculated in Eq.(1)

$$P(w^d = d | w^o = o, w^t = t; u) \propto P(w^d = d, w^o = o, w^t = t | u) \tag{3}$$

Case study and Results

We would like to have a ground truth to validate our algorithm, since successful inference means nothing to the credibility. Therefore, we use the smart card data of Metro system in Guangzhou—which records both boarding and alighting data—to perform the case study. There are total 170 stations and the time scope of the data is three month from July 1^{st} to September 30^{th} of year 2017. The operation time of the Metro system is 19 hours from 5:00 to 24:00, we set each hours as a temporal "word". Weekday and weekend are also be distinguished. Therefore, there are $19 \times 2 = 38$ kinds of trip departure time.

Among those with more than 20 records in the three month, We randomly select all the trip records of 2000 passengers. The total number of selected trips is 154276, which means averagely each person took 77 trips in the three month. Next, we randomly take off the destinations of 30% data and to see if the model can correctly infer the destinations by using the rest 70% data.

The number of latent topic should first be determined. Our tests show that the spatial dimension needs more latent topics than the temporal dimension, which is intuitively understandable. As shown in Table 1, further test shows that the number of destination topic L is a crucial parameter affecting the prediction accuracy. L need to be very large (compared to 170 stations) to improve the model performance, which is a drawback of current model.

It can been seen from Table 1 that the prediction accuracy does not improve significantly when the $L >= 130$. We adopt $J = 5, K = 10, L = 130$ and analyze how the number of training trips affects prediction performance, as shown in Figure 2. It can be found that the prediction accuracy is not high for less-frequent and over-frequent travellers. For the interval between 50 to 150 (corresponding to commuters), the model has the average accuracy around 60% to 80%.

Table 1: The prediction accuracy under different number of spatial latent topics when $J = 5$

number of origin topic K	Number of destination topic L						
	10	30	50	80	100	130	150
10	0.131	0.23	0.36	0.497	0.551	0.599	0.608
30	0.133	0.231	0.369	–	–	–	–
50	0.128	–	–	–	–	–	–
80	0.125	–	–	–	–	–	–

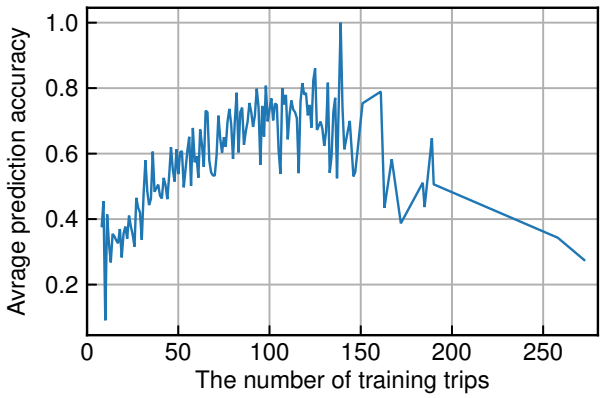

Figure 2: The average prediction accuracy for individuals with different number of training trips (When $J = 5, K = 10, L = 130$)

Discussion

Although the LDA model does not show advantage in terms of accuracy compared to existing destination estimation algorithm, the proposed model provides a whole new approach to address this problem. Besides making prediction, the latent features in the LDA model capture essential spatio-temporal characters of individuals and can be used of clustering and classification. The generative process also enables to generate synthetic data for activity-based simulation. Current model requires a very large destination topic number which significantly increase the computational burden; this problems should be solved by developing more proper structure for the generative process in future research.

References

J. J. Barry, R. Newhouser, A. Rahbee, and S. Sayeda. Origin and destination estimation in new york city with automated fare system data. *Transportation Research Record*, 1817(1):183–187, 2002.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022, 2003.

Z. Fan, A. Arai, X. Song, A. Witayangkurn, H. Kanasugi, and R. Shibasaki. A collaborative filtering approach to citywide human mobility completion from sparse call records. In *IJCAI*, pages 2500–2506, 2016.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101 (suppl 1):5228–5235, 2004.

S. Hasan and S. V. Ukkusuri. Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies*, 44:363–381, 2014.

L. He and M. Trépanier. Estimating the destination of unlinked trips in transit smart card fare data. *Transportation Research Record: Journal of the Transportation Research Board*, (2535):97–104, 2015.

M. A. Munizaga and C. Palma. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. *Transportation Research Part C: Emerging Technologies*, 24: 9–18, 2012.

A. A. Nunes, T. G. Dias, and J. F. e Cunha. Passenger journey destination estimation from automated fare collection system data using spatial validation. *IEEE transactions on intelligent transportation systems*, 17(1):133–142, 2016.

M.-P. Pelletier, M. Trépanier, and C. Morency. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568, 2011.

L. Sun, X. Chen, Z. He, and L. F. Miranda-Moreno. Pattern discovery and anomaly detection of individual travel behavior using license plate recognition data. In *Transportation Research Board 98th Annual Meeting*, 2019.

M. Trépanier, N. Tranchant, and R. Chapleau. Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, 11(1):1–14, 2007.