Feature Extraction for Long-term Travel Pattern Analysis

Yang Zhang[*1], Tao Cheng[†1]

[1]SpaceTimeLab for Big Data Analytics, Department of Civil, Environmental & Geomatic Engineering, University College London

January 12, 2017

Summary

Millions of urban dwellers with widely different travel patterns in London need to use public transit for daily travel. Most of existing travel-pattern analysis is short-term, based upon a short-period (one week or one month) of travel smart card (such as Oyster card in London) travel information. However, short-term results may vary according to different time periods or period length. For example, some passengers who only travel a lot in a short period should be treated as occasional users on an annul basis rather than regular users on a weekly basis. With respect to long-term transport planning and network development, a better understanding of passengers' long-term travel behaviour is important for transport agency. The main objective of this research is to give an explicit statement of feature extraction for long-term travel pattern identification. In this study, 25 distinct features are extracted from London Oyster card data to measure the spatial, temporal, and mode choice behaviours, which could give us a better understanding of the long-term travel patterns.

KEYWORDS: Oyster card, feature extraction, long-term travel pattern

1. Introduction

With millions of passengers using London public transport (PT) system every day, it is significant to better understand the passengers' travel patterns to improve the service quality of Transport for London (TfL). However, it is difficult to manually collate such information directly from the millions of PT users. From 2003, Oyster cards, a kind of contactless smart card, have been used on PT in Greater London to facilitate people's daily travel. After years of use, Oyster cards have amassed large amounts of transaction records, which provide an opportunity to investigate the passengers' travel patterns in Greater London.

Travel pattern identification usually is based on activity sequence (Lee and Park, 2005, Goulet Langlois et al., 2016) or extracted feature vectors (Ma et al., 2013). However, the dimension of activity sequence increases with the selected research duration, which limits its performance for long-term travel pattern recognition. The latter method can overcome this shortage.

Although feature extraction is a critical process of taking full advantage of original data to make clustering algorithms work, it has not been stated explicitly in current research, especially for long-term travel pattern analysis. To make up for this gap, in this study, 25 features related to passengers' spatial and temporal characteristics are extracted from Oyster card data to build the profiles. The details of feature definition, explanation and extraction process are given in this paper.

2. Methodology

[*] yang.zhang.16@ucl.ac.uk
[†] tao.cheng@ucl.ac.uk

2.1. Data description

The Oyster card data used in this research is a partial sample of the whole Oyster card transaction records in 2012 provided by TfL. Each transaction is collected automatically when a passenger taps in or out at a tube station or boards at a bus stop. Summarily, the entire dataset contains around 2.18 million journeys made by 9708 passengers, made up of 33.7% tube journeys and 66.3% bus journeys. In Oyster card data, each record contains an unique card ID, transport mode (bus or tube), transaction date, and entry/exist time and stations of a trip. However, the fare of a bus trip does not depend on the travel distance or zone. Thus, the alighting time and station of a bus trip are never recorded.

2.2. Feature extraction

Establishing the users' profiles is critical to segment passengers for travel pattern analysis. In some existing works, travel pattern recognition based on feature extraction has been studied. For example, Ortega-Tong (2013) defined 20 features in spatial, temporal or economic dimensions to classify London's PT users into 8 groups for travel pattern investigation by using one-week Oyster card data. However, the author ignored some indicators for long-term travel pattern identification, such as the period of using Oyster card. In addition, the economic-related features of some passengers are ambiguous because of the complex fare structure. For instance, Oyster card users who hold Travelcard (a weekly, monthly or annual pass ticket) can have unlimited tube trips within the purchased zones and bus trips of all bus lines, then the fare of a single trip is hard to define explicitly. Furthermore, some geographical information is not considered, such as the travel zone, which can reflect a passenger's activity area.

To understand who the passenger is and why the passenger travels, we consider the travel features could be categorised as three categories: temporal variability (when the passenger travels), spatial variability (where the passenger travels), and travel mode preference (how the passenger travels). Note that, economic-related features are not considered as an extra category, because sometimes the fare of a single trip is unavailable. For example, if a card holder buys a monthly Travelcard (a pass ticket), we cannot define the cost of each trip within the purchased month. To avoid the ambiguous results, economic-related features are excluded. However, the ticket price is related to the travel zone, length, time and frequency, and all these characteristics will be considered in the three categories inexplicitly.

Feature extraction process complies with the rules of simplicity, non-redundancy and discriminative ability. In this study, 25 features are extracted to differentiate the distinct travel behaviours, as shown in Table 1.

In temporal features, the six features measure the travel time distribution of a passenger within a day. AFTI_WD and LFTI_WD may indicate the working status or travel purpose on weekdays while the AFTI_WE and LFTI_WE indicate the regularity of leisure behaviours. MPT_NUM and EPT_ NUM are extracted considering two peak times defined by TfL: 7:00AM-10:00AM and 4:00PM-7:00PM from Monday to Friday. These two factors cannot only reveal the temporal information, but also reflect economic characteristics to some extent, because the ticket prices during peak hours are higher than during off-peak. The last three temporal features are related to travel frequency, indicating the passengers' travel regularity. ACTI_DUR, an important indicator for long-term pattern analysis, is computed by the last travel date minus the first one in year 2012. For a long-term travel pattern analysis, we cannot only care about how many days a passenger travels (ACTI_DAY), but also the duration of travel activities, because the two factors can imply the travel density (ACTI_DAY/ ACTI_DUR) through PT.

Table 1 Extracted features for passenger profiling

Subgroup	Feature	Description
Temporal features	AFTI_WD	The average first tap_in time on weekdays
	LFTI_WD	The average last tap_in time on weekdays
	AFTI_WE	The average first tap_in time on weekends
	LFTI_WE	The average last tap_in time on weekends
	MPT_NUM	the number of trips during morning peak (7:00am-10:00am)
	EPT_ NUM	the number of trips during evening peak (4:00pm-7:00pm)
	AVG_TRIP	The average number of trips per day
	TRA_DAY	How many days a passenger travels in the whole year
	ACTI_DUR	Active duration in the whole year
Spatial features	AVG_ TIME_WD	The average of tube trip time on weekdays
	VAR_ TIME_WD	The variance of tube trip time on weekdays
	AVG_ TIME_WE	The average of tube trip time on weekdays
	VAR_ TIME_WE	The variance of tube trip time on weekdays
	MAX_TD	The average radius travelled by tube per day
	AVG_TS	The average of the number of different tube stations used per day
	VAR_TS	The variance of the number of different tube stations used per day
	AVG_BL	The average of the number of different bus lines used per day
	VAR_BL	The variance of the number of different bus lines used per day
	AVG_INNER	The mean value of the inner zone number
	AVG_OUTER	The mean value of the outer zone number
Mode preference	TUBE_NUM_WD	The total number of the tube journeys on weekdays
	BUS_NUM_WD	The total number of the bus journey on weekdays
	TUBE_NUM_WE	The total number of the tube journeys on weekends
	BUS_NUM_WE	The total number of the bus journey on weekends
	MODE_T	How often a passenger change the transport mode per day? (average)

Spatial features can provide spatial behaviour information and geographical information. The first four features are about the travel time spent on the transport network, which imply the real commuting distance and the travel variability. MAX_TD is the mean value of travel radius calculated only by using tube journeys, because of the missing alighting stations of bus trips. Travel radius means the maximum Euclidean distance between any two different tube stations used in a day, indicating the range of activity. The next four features (AVG_TS, VAR_TS, AVG_BL, and VAR_BL) can describe the travel regularity of a card holder. Since the alighting bus stops are not available, we use the bus line to indicate the regularity of the travel by bus. The last two features in this group capture the geographical information related to tube journeys. Tube services in London are divided into 1-9 zones (ring shape) while buses operate regardless of the zones. AVG_INNER and AVG_OUTER cannot only reveal the passengers' activity location, but also reflect the economic-related features, because the ticket prices can vary considerably according to which and how many zones a passenger travels through.

The last group is mode-preference features. TUBE_NUM_WD, BUS_NUM_WD, TUBE_NUM_WE, and BUS_NUM_WE reflect the PT mode choice of a passenger, which is related to travel purpose, economic status and so on. For example, commuters having fixed working hours may choose to take tube, because tube trip time is usually shorter than bus trip and it is seldom influenced by traffic congestion. However, the bus ticket price is higher than tube's, thus it can reflect the economic status of a passengers. Then, we use another indicator, MODE_T, to measure how often a passenger change the transport mode per day. MODE_T can indicate the convenience of PT and the accessibility from an origin to a destination.

These 25 descriptive variables can characterise passengers' distinct travel behaviour from temporal, spatial and travel mode perspectives. Before using these extracted features to classify passengers' travel patterns, all features are standardised between 0 and 1. This step can eliminate the influence of

magnitude and variability of extracted features, and keep all information of original. Then, passengers can be segmented by using the 25-dimension feature vector to identify their diverse travel patterns.

3. Summary and future work

In this research, a detailed and explicit statement of feature extraction from Oyster card data for long-term travel pattern analysis is illustrated. The 25 variables extracted from transaction records are categorized into 3 groups, describing temporal, spatial and PT mode preference features. Future work will focus on travel pattern classification, validation and social-demographic analysis.

4. Acknowledgements

The first author's PhD research is jointly funded by China Scholarship Council and the Dean's Prize from the University College London. The data provided by Transport for London (TfL) is highly appreciated.

5. Biography

Yang Zhang is a PhD student in department of Civil, Environmental and Geomatic Engineering at University College London. Her research interest includes spatial-temporal data analytics, smart card data mining and complex. Her current work is focus on passenger profiling and travel pattern recognition.

Tao Cheng is a Professor in GeoInformatics, and Director of SpceTimeLab for Big Data Analytics (http://www.ucl.ac.uk/spacetimelab), at University College London. Her research interests span network complexity, Geocomputation, integrated spatio-temporal analytics and big data mining (modelling, prediction, clustering, and simulation), with applications in transport, crime, health, social media, and environmental monitoring.

References

GOULET LANGLOIS, G., KOUTSOPOULOS, H. N. & ZHAO, J. 2016. Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies,* 64**,** 1-16.

LEE, K. & PARK, J. S. Traversal pattern analysis of transit users in the Metropolitan Seoul. Proceedings of International Forum on the Public Transportation Reform in Seoul, 2005. 7-8.

MA, X., WU, Y.-J., WANG, Y., CHEN, F. & LIU, J. 2013. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies,* 36**,** 1-12.

ORTEGA-TONG, M. A. 2013. *Classification of London's public transport users using smart card data.* Massachusetts Institute of Technology.