Transportation Research Part C 46 (2014) 179–191

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Behavioural data mining of transit smart card data: A data fusion approach

Takahiko Kusakabe *, Yasuo Asakura

Department of Civil Engineering, Tokyo Institute of Technology, Japan

ARTICLE INFO

Article history:
Received 8 November 2013
Received in revised form 22 May 2014
Accepted 22 May 2014

Keywords:
Smart card data
Data fusion
Data mining
Behavioural analysis
Naïve Bayes classifier

ABSTRACT

The aim of this study is to develop a data fusion methodology for estimating behavioural attributes of trips using smart card data to observe continuous long-term changes in the attributes of trips. The method is intended to enhance understanding of travellers' behaviour during monitoring the smart card data. In order to supplement absent behavioural attributes in the smart card data, this study developed a data fusion methodology of smart card data with the person trip survey data with the naïve Bayes probabilistic model. A model for estimating the trip purpose is derived from the person trip survey data. By using the model, trip purposes are estimated as supplementary behavioural attributes of the trips observed in the smart card data. The validation analysis showed that the proposed method successfully estimated the trip purposes in 86.2% of the validation data. The empirical data mining analysis showed that the proposed methodology can be applied to find and interpret the behavioural features observed in the smart card data which had been difficult to obtain from each independent dataset.

1. Introduction

Smart card systems have been installed as a method to collect fare of public transport. These systems automatically and continuously collect the records of travellers' use of the public transport because records of the fare payment can be regarded as travel records. For example, the transaction data enable us to observe volume of passengers at the point of ticket gates. That is; Smart card systems collect ID information of each traveller's smart card alongside fare collection. IDs enable us to analyse travel patterns such as each traveller's trip frequency, travel sections, and trip sequences. Travel patterns and their variability over long-term periods can thus be analysed (e.g. Bagchi and White, 2005 and Utsunomiya et al., 2006). Therefore, ID information of each traveller and long-term continuous observation are the advantages of the smart card data comparing to other conventional data.

Several studies have been attempted to develop methods to apply smart card data to analyses for transit management and planning. Chu and Chapleau (2008) presented methods to estimate the arrival times of a passenger at a bus stop and to identify linked trips using spatial–temporal concepts. Trépanier et al. (2007) presented a model to estimate the destination location for each individual boarding by using the smart card data. Seaborn et al. (2009) analysed multimodal journeys for information on transit planning using the smart card system in London by identifying the multimodal transfer combinations of bus-to-Underground, Underground-to-bus, and bus-to-bus. Kusakabe et al. (2010) and Asakura et al. (2012) estimated boarding trains of railway passengers by using smart card data and evaluated the effects of changes in train

* Corresponding author. Address: 2-12-1-M1-20, O-okayama, Meguro, Tokyo 152-8552, Japan. Tel./fax: +81 3 5734 2575.
E-mail address: t.kusakabe@plan.cv.titech.ac.jp (T. Kusakabe).

http://dx.doi.org/10.1016/j.trc.2014.05.012
0968-090X/

operations. They compared passengers' travel choice behaviour before and after the railway company altered the train timetable. Ma et al. (2013) developed an efficient data mining method to demonstrate the temporal travel patterns and the pattern regularity for transit riders in Beijing. Pelletier et al. (2011) categorised the previous studies on smart cards in public transport. They classified the usage of data into three purposes: strategic (long-term planning, behavioural analysis, and demand forecasting), tactical (service adjustments and network development), and operational (ridership statistics and performance indicators) purposes.

Continuous observations on travel patterns are important for transport operators to assess the current state of the effects of their efforts, such as timetable improvements and transit planning. Previously, the methods to collect the data on transit behaviour have mainly relied on behavioural surveys. The conventional behavioural survey methods are specialised in collecting data items for behavioural analysis and transit planning. For example, household surveys such as person trip surveys are used to gather data on all trips throughout a day with trip purposes, travel modes, actual origins, and destinations. In order to obtain dynamic changes in the travel behaviour, previous studies have developed advanced survey methods such as travel diary surveys (e.g. Pas and Koppelman, 1986; Axhausen et al., 2002), panel surveys (e.g. Kitamura, 1990), and travel surveys with tracking devices (e.g. Murakami and Wagner, 1999 and Asakura and Hato, 2004). They showed methods using the dynamic travel demand observation to identify temporal variation in travel behaviour, and to evaluate the impact of a change in the transportation system. However, continuous long-term observation is still difficult by using these specially designed surveys. Kitamura (1990) indicated that the disadvantages of a panel survey are possible increase in non-responses, problem of attrition (a decrease in the number of panels between the waves of the survey), possible decline in reporting accuracy due to panel fatigue, and problem of panel conditioning. Previous studies (e.g. Golob and Meurs, 1986; Kitamura and Bovy, 1987; Van Wissen and Meurs, 1989) pointed out that the number of responses by each respondent declines gradually during the survey period. These disadvantages are possibly found in travel data designed by other methods when the surveys are conducted for long-time period. To prevent such declines, long-term great efforts by survey staffs such as interviewers are needed to collect sufficient number of data (e.g. Axhausen et al., 2007). Although the information technologies enable us to automatically track the respondents where number of respondents in the most of tracking surveys remains less than a thousand, and duration of the tracking travel surveys is less than a few months (e.g. Murakami and Wagner, 1999; Asakura and Hato, 2004; Wolf et al., 2001, and Draijer et al., 2000). It is still difficult to collect day-to-day data for continuous long-term periods via the behavioural surveys because of cost, processing load, accuracy, and privacy protection of respondents. It means that survey based data are practically not available for monitoring long-term characteristics of transport demand. Behavioural surveys are more suitable for observing some specific factors to implement the planning and management policies rather than for continuous monitoring of travel demand.

The smart card data provide continuous and long-term travel information. However, they are fragmentary for behavioural analysis. For example, the data do not include the travellers' origins, destinations or trip purposes, or the data do not cover travellers' behaviour throughout the entire transport network. This is because the data are not collected with an explicit goal of behavioural analysis; rather, they are happened to be collected while fare collection. Trépanier et al. (2009) compared household travel survey data with smart card data. They showed that the large variations in transit network use on weekdays can be captured by using smart card data, although the smart card data are partially consistent with the travel survey data. Their result implies that the data is applicable to analyse transit behaviour if the insufficient parts of the data are supplemented or negligible. Several studies have attempted to develop methods to obtain the user segments and their behavioural contexts from behavioural patterns observed in smart card data. Agard et al. (2006) and Morency et al. (2007) estimated behavioural pattern groups and showed the variability of travellers' behavioural patterns. Kusakabe and Asakura (2011) proposed a method to identify within-day and day-to-day behavioural patterns of smart card users by a latent class model. However, the meanings of the segments, which are related to their activities, should be subjectively interpreted by analysts in these studies.

Data fusion is one of the approaches to integrate multiple data sources, which is applied in various fields, such as military applications, marketing, and intelligent transportation systems (e.g. Hall, 1992; Mitchell, 2007; Kamakura and Wedel, 1997; El Faouzi et al., 2011, and Shen and Stopher, 2013). For example, Shen and Stopher (2013) developed a trip purpose imputation method for Global Positioning System (GPS) data by using the National Household Travel Survey (NHTS) in the US. In their method, the trip purposes which were not directly observed by GPS data were estimated using rules obtained from the NHTS data. This study employs the data fusion methodology to derive relationships among behavioural attributes that cannot be obtained from either smart card data or survey based data alone. The survey based data directly observe detailed information on travel behaviour but being unable to continuously do so over a long-term period. In contrast, the smart card data provide only fragmentary information on travellers' behaviour though they can provide a continuous long-term period data which is difficult to achieve via a person trip survey. If the advantages of smart card data are combined with that of the person trip survey data, it would improve the effects of continuous monitoring of transport demands. Therefore, data fusion will allow us to obtain good understanding of the changes in travellers' behaviour over the continuous long-term period.

The aim of this study is to develop a data fusion methodology to estimate absent behavioural attributes in smart card data by using survey based data. The proposed method intends to enhance understanding of travellers' behaviour during monitoring of the smart card data. The method illustrates the relationship between original observed attributes of smart card data and the estimated attributes that are unobservable. In order to utilise the smart card data with survey based data, this study applies the naïve Bayes classifier method. A person trip survey data is employed as a survey based data to estimate the probability distribution of the naïve Bayes probabilistic model.