Transportation Research Part C 63 (2016) 170–181

Contents lists available at ScienceDirect

Transportation Research Part C

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

Mining missing train logs from Smart Card data

Yun-Hong Min [1], Suk-Joon Ko *, Kyung Min Kim [2], Sung-Pil Hong

Dept of Industrial Eng., Seoul National University, San 56-1 Shilim-dong, Kwanahk-gu, Seoul 151-742, Republic of Korea

ARTICLE INFO

Article history:
Received 3 June 2015
Received in revised form 28 November 2015
Accepted 30 November 2015
Available online 2 January 2016

Keywords:
Data mining
Smart Card data
Train log
Passenger behavior

ABSTRACT

This paper shows how to recover the arrival times of trains from the gate times of metro passengers from Smart Card data. Such technique is essential when a *log*, the set of records indicating the actual arrival and departure time of each bus or train at each station and also a critical component in reliability analysis of a transportation system, is missing partially or entirely. The procedure reconstructs each train as a sequence of the earliest exit times, called *S-epochs*, among its alighting passengers at each stations. The procedure first constructs a set of passengers, also known as *reference passengers*, whose routing choices are easily identifiable. The procedure then computes, from the exit times of the reference passengers, a set of tentative *S*-epochs based on a detection measure whose validity relies on an extreme-value characteristic of the platform-to-gate movement of alighting passengers. The tentative *S*-epochs are then finalized to be a true one, or rejected, based on their consistencies with bounds and/or interpolation from prescribed *S*-epochs of adjacent trains and stations. Tested on 12 daily sets of trains, with varying degrees of missing logs, from three entire metro lines, the method restored the arrival times of 95% of trains within the error of 24 s even when 100% of logs was missing. The mining procedure can also be applied to trains operating under special strategies such as short-turning and skip-stop. The recovered log seems precise enough for the current reliability analysis performed by the city of Seoul.

1. Introduction

Buses and trains are dictated by operation plans, including schedule timetables and headways, on which each passenger relies to plan his or her route. Yet road congestions, increased dwell times, accidents, and other factors can all distort a planned operation and lead to degraded reliability of public transportation services. As a result, we often find ourselves stranded on a station or inside a packed car.

Transit service reliability is defined as, according to KFH Group (2013), "how often service is provided when promised." and is measured by on-time performance, headway adherence, missed trips, and distance traveled between mechanical breakdowns. Hence the availability of a *log*, i.e. the set of records indicating the actual arrival and departure time of each bus or train at each station, is critical for reliability analysis.

Introduction of automatic passenger counters (APC) and automatic vehicle locators (AVL) for buses, and the control system for trains including automatic train operation (ATO) have made available precise logs of public transport. Hammerle

* Corresponding author. Tel.: +82 2 884 1164.
E-mail address: reclus85@snu.ac.kr (S.-J. Ko).
[1] Current address: Software Solution Laboratory, Samsung Advanced Institute of Technology, Republic of Korea.
[2] Current address: Policy-Technology Convergence Research Division, Korea Railroad Research Institute, Republic of Korea.

http://dx.doi.org/10.1016/j.trc.2015.11.015
0968-090X/

et al. (2005) and El-Geneidy et al. (2011), for instance, employed logs obtained from APC and AVL in evaluating on-time performance and headway adherence of buses in Chicago and Minneapolis, respectively. Logs, however, are not always available or, if so, complete. New York City transit lacked any system to acquire logs of their vehicles in 2000 (Doyle, 2000). Zurich, though having installed AVLs on all of their buses, had only 10% of them equipped with APCs by 2009 (Orth et al., 2012). Some of Seoul's early metro stations kept no logs; others archived them incompletely or on unstable form, such as magnetic tapes, that later deteriorated beyond repair.

Fig. 1 is the metro network of Seoul metropolitan area as of November 2011. Black region includes 245 stations that maintain complete logs. The logs of remaining 248 stations are missing partially or entirely. Only 6 of 15 lines maintain complete logs for each train they operate.

This paper aims at recovering missing logs from the Smart Card Automated Fare Collection System, or *Smart Card*, data. In Seoul, after 9 years of practice, Smart Card has become the sole method of payment for the city's metro as of 2009. Due to its massive data and comprehensive categories, many studies have been conducted to fully realize the potential it holds (Pelletier et al., 2011); *O–D* matrix estimation (Cui, 2006; Trépanier et al., 2007; Munizaga and Palma, 2012), route choice estimation (Kusakabe et al., 2010; Hong et al., 2015), transit behavior analysis (Lathia and Capra, 2011; Ma et al., 2013; Kim et al., 2015), and demand-driven timetabling (Niu and Zhou, 2013; Sun et al., 2014) are few of many examples.

Smart Card data includes the *quadruples*, (Departure station *O*, Entry time at gate, Arrival station *D*, Exit time at gate), in addition to vehicle ID, fare, distance traveled and seniority/disability status of the card holder. In this paper, we only use the set of quadruples, which is the minimum data needed to reliably reconstruct logs as well as the maximum we can expect to obtain from any metro network employing a Smart Card system.

Our procedure reconstructs each train as a sequence of the earliest exit times, called *S-epochs,* among its alighting passengers at each station. Thus we need the exit times of alighting passengers assorted to their trains at every station. However, the exit times from Smart Card data are recorded tag times of passengers in their order of arrivals at a gate often shared by trains from different directions – in and outbound – and/or even from different lines at a transfer station. Hence, in the first step, we extract from Smart Card data a considerable set of passengers whose routing choices are easily traced from their origin and destination pairs. Such passengers, known as *reference passengers*, give us, with certainty, the directions of their trains and the lines to which they belong, if not the exact trains themselves.

Based on the exit times of reference passengers, the procedure then computes a set of tentative *S*-epochs based on a detection measure whose validity relies on an extreme-value characteristic of the platform-to-gate movement of alighting passengers; the moments of exit at gates are closely grouped for passengers from identical trains. The tentative *S*-epochs are then finalized to be a true one, or rejected, based on their consistencies with bounds and/or interpolation from prescribed *S*-epochs of adjacent trains and stations. The arrival times of trains can then be recovered by shifting the finalized *S*-epochs backward in time by the platform-to-gate time of each station.

Tested on the instances from three entire metro lines from Seoul metropolitan area with varying degrees of missing logs, the mining procedure computed the arrival times of 95% of trains within the error of 24 s for local-only lines and within 45 s for lines operating local and express trains even when the logs were missing for an entire line. In the reliability analysis currently practiced by the city of Seoul, a train is defined to be delayed if it is late by more than 10 min. Therefore, our level of error, less than 8% of the threshold at maximum, seems permissible for the recovered log to replace the real one.

This paper is organized as follows. Section 2 describes the mining procedure, *S*-epochs, reference passengers, extreme-valued nature of the platform-to-gate times of metro passengers that enables the detection measure of *S*-epochs, and consistency checks that finalize the tentative *S*-epochs along with their applications on real instance. Section 3 describes how the procedure can be extended to mining trains under special operation strategies including short-turning and skip-stop. Section 4 evaluates the performance, especially the accuracy of the procedure. Finally, Section 5 provides some concluding remarks and possible further research.

2. Mining procedure

Our objective is to recover missing arrival times of a train in a given metro line using the quadruples – (Departure station *O*, Entry time at gate, Arrival station *D*, Exit time at gate) – of metro passengers obtained from Smart Card data. The proposed mining procedure first recovers each missing train *X* as the sequence of *S*-epochs $s_{X,N}$, the earliest exit time of a passenger at a gate of stations *N* at which *X* stops, which is then shifted backward in time into arrival times. The main steps of the procedure are pre-summarized in Fig. 2.

The proposed method is applicable, in principle, when logs are missing for an entire line. In this section, for simplicity, we first restrict the discussion to the case where logs are available for a subset of stations. Then, in Section 3, we discuss how the method can be extended to recover the trains when logs are entirely missing.

2.1. S-epochs

The upper graph in Fig. 3 plots the entry times of the 95 passengers traveling downward from Moran to Seohyeon on Bundang Line of Seoul metro that entered the platform of origin, Moran, between 7:45 and 8:03 A.M. on November 21, 2011. As expected, passengers arrive onto the platform randomly over time without any noticeable pattern.