Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 52 (2015) 67 – 74

Procedia Computer Science

The 6th International Conference on Ambient Systems, Networks and Technologies (ANT 2015)

Deriving public transportation timetables with large-scale cell phone data

Christopher Horn, Roman Kern

Know-Center Gmbh, 8010 Graz, Austria

Abstract

In this paper, we propose an approach to deriving public transportation timetables of a region (i.e. country) based on (i) large-scale, non-GPS cell phone data and (ii) a dataset containing geographic information of public transportation stations. The presented algorithm is designed to work with movements data, which are scarce and have a low spatial accuracy but exists in vast amounts (large-scale). Since only aggregated statistics are used, our algorithm copes well with anonymized data. Our evaluation shows that 89% of the departure times of popular train connections are correctly recalled with an allowed deviation of 5 minutes. The timetable can be used as feature for transportation mode detection to separate public from private transport when no public timetable is available.

Peer-review under responsibility of the Conference Program Chairs

Keywords: Cell phone data; Transportation Mode Detection; Public Transportation;Timetable;

1. Introduction

When conducting transportation mode detection for given movement trajectories, it is necessary to use features that are as expressive as possible. For example, Stenneth et al.[1] used the distance to the next bus stop location as a feature for their classifier. Another feature that may help to distinguish between public and individual transportation is based on public transportation timetables: the closer the travel time approximates the departure and arrival times of a public transportation facility, the more likely its trajectory is to originate from this facility. However, it might not be possible to obtain the official timetable of a region by automatic means. In the paper, we propose an approach to inferring the train timetable of a given region from movement-based datasets and locations of train stations.

Our dataset consists of anonymised cell phone events that are based on interactions with a cell phone providers infrastructure rather than on GPS data or sensor data directly recorded on the cell phone. Since such events are

* Corresponding author. Tel.: +43 316 810-30866 ;
E-mail address: chorn@know-center.at

Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2015.05.026

infrequent and have a low spatial accuracy, this task is particularly challenging. The discrepancy between the observed location and the true location can be two orders of magnitude bigger than in existing GPS-based systems. Depending on the users behaviour, the frequency may vary from a single event every 6 hours to several events per minute. The cell phone data was pre-processed to avoid the identification of individual users by removing all personal information from the dataset and replacing the phone number with a hash value that was valid for 24 hours. This allowed us to establish anonymous movement trajectories for up to 24 hours. In addition, to prevent the exposure of individual users and their positions, the dataset included obfuscated and spurious dummy events.

Although our dataset and task were quite specific, the results should apply to a much bigger class of location- and movement-aware data analytics scenarios. For example, they may offer insights into the service-privacy trade-off that applies to many location-based services, some of which deliberately cloak the users privacy profile and insert dummy trajectories (e.g., k-anonymity algorithms and path confusion approaches, such as Never Walk Alone). Gruteser and Grunwald[2] proposed various algorithms that meet certain anonymity requirements by decreasing the spatial or the temporal resolutions. Our work demonstrates the usefulness of even highly obfuscated movement datasets. In our evaluation, we provide a deep analysis not only of the individual parameters of our algorithm, but also measure how much input data is required to achieve a certain accuracy.

2. Related work

Extensive research of transportation mode detection was performed using high-sampling GPS[1,3,4,5] or GPS data in connection with on-device sensors, such as an accelerometer[6]. In our work, we employ low-sampling and low-accuracy cell phone data.

Wang et al.[7] used CDRs to infer the transportation mode based on the travel time. They used a k-Nearest-Neighbor (kNN)-based clustering approach to distinguish between the different types of transportation modes (cars, public transportation, walking). In our setting, the travel times of car and train travellers partly overlapped, making a distinction based solely on the travel time unfeasible. Sohn et al.[8] applied manually collected GSM traces to distinguish between the three mobility modes (“stationary”, “walking” and “driving”). Unlike other approaches, this one relies on the signal strength and the change between two consecutive measurements rather than on the geographic coordinates of the cell towers.

In the field of public transportation modelling, Aguilera et al.[9] used cell phone data to measure passenger flows in the Paris transit system. In addition to travel times, they derived occupancy rates and origin-destination flows and established that in 80% of cases the occupancy rates estimated by GSM data corresponded to the actual ones. Calabrese et al.[10] fused cell phone data with the location data of public transportation, which allowed the authorities to better understand the movement patterns of pedestrians and buses.

However, to the best of our knowledge, no research that uses large-scale cell phone data to derive a public transportation timetable has been performed to date.

3. Methodology

In this section we describe the proposed approach to deriving the public transportation timetable of a region (i.e., a country) based solely on map data and non-GPS cell phone data. The assumption was that it was possible to use a large-scale movement dataset to identify bursts of travellers that in turn indicate public transportation movements, even if the single trajectories failed to reliably detect such movements. This assumption is visualised in Figure 1 that shows the transitions between nearby checkpoints: Figure 1 (a) the checkpoints are two motorway junctions and Figure 1 (b) the checkpoints are two railway stations. Unlike the transitions between two motorway junctions, those between two railway stations indicate multiple bursts (spikes). Assuming that those bursts are moving trains, we can derive an accurate train timetable based on a burst detection algorithm. In the remainder of this paper, we demonstrate our method using train timetables as an example.

Obtaining a train timetable of a region required the following: (i) a dataset containing geolocations for public transportation, (ii) a large-scale movement dataset, (iii) an algorithm to calculate the transitions, and (iv) a burst detection algorithm to derive the timetable.