Computers, Environment and Urban Systems 34 (2010) 541–548

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Uncovering cabdrivers' behavior patterns from their digital traces

Liang Liu *, Clio Andris, Carlo Ratti

Department of Urban Planning and Studies, Massachusetts Institute of Technology, MIT, 10-485, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 8 December 2009
Received in revised form 11 July 2010
Accepted 17 July 2010

Keywords:
Cabdriver
Operation behavior
Mobility pattern
GPS
Spatial–temporal data mining
Pervasive computing

ABSTRACT

Recognizing high-level human behavior and decisions from their digital traces are critical issues in pervasive computing systems. In this paper, we develop a novel methodology to reveal cabdrivers' operation patterns by analyzing their continuous digital traces. For the first time, we systematically study large scale cabdrivers' behavior in a real and complex city context through their daily digital traces. We identify a set of valuable features, which are simple and effective to classify cabdrivers, delineate cabdrivers' operation patterns and compare the different cabdrivers' behavior. The methodology and steps could spatially and temporally quantify, visualize, and examine different cabdrivers' operation patterns. Drivers were categorized into top drivers and ordinary drivers by their daily income. We use the daily operations of 3000 cabdrivers in over 48 million of trips and 240 million kilometers to uncover: (1) spatial selection behavior, (2) context-aware spatio-temporal operation behavior, (3) route choice behavior, and (4) operation tactics. Though we focused on cabdriver operation patterns analysis from their digital traces, the methodology is a general empirical and analytical methodology for any GPS-like trace analysis. Our work demonstrates the great potential to utilize the massive pervasive data sets to understand human behavior and high-level intelligence.

Published by Elsevier Ltd.

1. Introduction

In recent years, the widespread deployment of pervasive computing technologies in cities has led to a massive increase in the volume of records of human spatio-temporal paths throughout the built environment. Collectively, these records allow for the development of new ways to study human behavior and provide new opportunities for the promising field of computational social science (Lazer et al., 2009).

Different types of data exist and can be used in the urban context. Some researchers have recently focused on the cellular network to study city dynamics and human mobility (González, Hidalgo, & Barabási, 2008; Ratti, Pulselli, Williams, & Frenchman, 2006; Reades, Calabrese, Sevtsuk, & Ratti, 2007). Others have been taking advantage of the data combinations produced by the overlay of digital information onto urban infrastructure – such as bus and subway routes, taxis, public utilities, and roads. Finally, much attention has been given to GPS-based data analysis, in order to understand individual outdoor movement, extract main destinations, predict movements and model routines (Ashbrook & Starner, 2003; Krumm & Horvitz, 2006; Liao, Fox, & Kautz, 2004; Liao, Patterson, Fox, & Kautz, 2005; Patterson, Liao, Fox, & Kautz, 2003). Many of these studies use aggregate data, and so far, few have used individual traces to better understand high-level human behavior and decisions, which would enable the creation of new services that autonomously respond to a person's unspoken needs.

This study aims to fill this gap in research through the analysis of an unprecedented database of individual, economically-driven spatio-temporal choices: a collection of records from 3000 cabdrivers, who take a total of 48 million trips covering 240 million kilometers for one year in Shenzhen, South China. We know the drivers are in competition to earn the most income via strategic operations patterns, but we are surprised to observe a noteworthy variation in 'skill level', as some drivers earn consistently up to six times more than others.

From this starting point, we refine our avenues of inquiry: What features of this very large database will lead to uncovering top drivers' strategy? How do these successful drivers optimize over the bounded resources of space and time? Or are these drivers profiting by deliberately choosing routes that are more costly for the customer, as every out-of-towner fears when stepping into a taxicab in an unfamiliar city? In solving these problems, we aim to develop a new methodology for the analysis of large amounts of spatio-temporal traces that, in the future, could be applied to other cities and domains.

Our contribution lies in the following:

(1) For the first time, we systematically study large scale cabdrivers' behavior in a real and complex city context (3000 taxis in a metropolitan area) through their daily digital traces. We identify a set of valuable features, which are

* Corresponding author. Tel.: +1 617 2535558; fax: +1 617 2588081.
E-mail addresses: liuliang@mit.edu (L. Liu), clio@mit.edu (C. Andris), ratti@mit.edu (C. Ratti).

0198-9715/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.compenvurbsys.2010.07.004

simple and effective to classify cabdrivers, delineate cabdrivers' operation patterns and compare the different cabdrivers' behavior.

(2) We develop a novel methodology and steps to spatially and temporally quantify, visualize, and examine cabdrivers' operation patterns. Drivers are categorized into top drivers and ordinary drivers by their daily income. We use the daily operations of 3000 cabdrivers in over 48 million of trips and 240 million kilometers to uncover the differences between top drivers and ordinary drivers: (1) spatial selection behavior, (2) context-aware spatio-temporal operation behavior, (3) route choice behavior, and (4) operation tactics.

The paper is structured as follows. First, we describe the raw data and our processing and feature extraction algorithm. Second, we propose a way to classify different cab drivers into high-earning 'top' drivers and average-earning 'middle' drivers (also referred to as average or ordinary) based on their income and rank. Third, we compare the different statistical operation patterns of top and middle drivers to reveal the "invisible hand" of competitive driving strategy. Finally, we discuss our results and possible directions for future work.

2. Data sets

2.1. Data description

We have continuously collected data from 3000 taxi drivers' GPS traces for over 1 year (2008), yielding accurate data about taxi location in longitude, latitude form, timestamp, vehicle identification, operation status (empty or occupied), spot speed, and azimuth. The GPS traces include both occupied and empty running trips in order to describe the full operation behavior of cabdrivers. We also have contextual traffic condition information (car speed and volume), collected by the Shenzhen Transport Information Center and based on floating car and fixed point sensors.[1]

Some information should be given about the *modus operandi* of taxicab drivers in Shenzhen. First, they do not work for a taxi dispatching system but simply for the hail market, meaning they cruise the road to find passengers. Most cars run 24 h a day in two shifts: the first shift (day) is from 7 am to 7 pm, the other (night) from 7 pm to 7 am. A single pair of drivers is linked to each vehicle, so we can actually consider each cab as a 2-person contained unit.

We should also mention that drivers do not have any significant spatial limitations; instead, they can operate freely in any part of the city and can choose their working zones. Finally, the fare structure for taxi drivers in Shenzhen is the following: 12.5 Yuan for the first 3 km and the 2.4 Yuan charge for each additional kilometer. This fare system varies in different parts of the world, and Shenzhen's system will certainly play a role in temporal operation decisions.

2.2. Initial data processing and feature extraction

After extracting important features from the GPS traces, we perform data cleaning, coordinate transformation, and map matching processes in order to understand cabdrivers' choices. In particular:

(1) Trip distance, travel time, travel speed, income/trip. For every trip, including occupied and empty legs, we calculated the trip distance on the road network, the corresponding travel time, travel speed and income. Of course, for trip fragments without a passenger, the income is zero.

(2) Origin/destination pair of each trip. For each trip, we estimate the origin (pick up location) and destination (drop off location) of the trip, hence gaining a good understanding of taxi travel demand at different times of the day. We consider origin to be a transition from empty to occupied in our timestamp sequence, and a destination to be a transition from occupied to empty in our sequence.

(3) Ratio of real path length over shortest path length (RRSL) and ratio of real path travel time over shortest path travel time (RRST). For each trip, given specific origin and destination, we calculate the shortest path length on the road networks and get the ratio of real trip distance over shortest path length (RRSL). Turner (2009) recently referred to this ratio as 'angularity' when analyzing courier GPS traces. At the same time, we also estimate the travel time, given the real time traffic speed on the road links if the driver uses the shortest paths, then we calculate the ratio of real path travel time over shortest path travel time (RRST).

Jiang, Yin, and Zhao (2009) look at taxi cab travel on a large road network over 6 months. In comparison, they simulate a Levy Flight process, where the of steps in a walk are chosen randomly from a fat tailed distribution of directions (e.g. left, right, up, and down) and mimic the degrees of freedom provided by the road length infrastructure. Their findings show that real data from 50 taxis exhibit this flight behavior because of the constraints of the underlying street network.

3. Classification of taxi drivers

3.1. Calculation of taxi income

Our first analysis is based on the calculation of the daily income of taxi drivers. We derive this measure from the fare structure discussed in Section 2.1, and our calculations of trip length on the road grid, and vehicle status (occupied/empty). We find that the average daily income follows a normal distribution for all taxi drivers for 1 year (Fig. 1).

Most drivers (ordinary drivers) garner income are around the mean value (700–800 Yuan, approximately 110 US$), several are very high (top drivers, 1250 Yuan), several very low (200–300

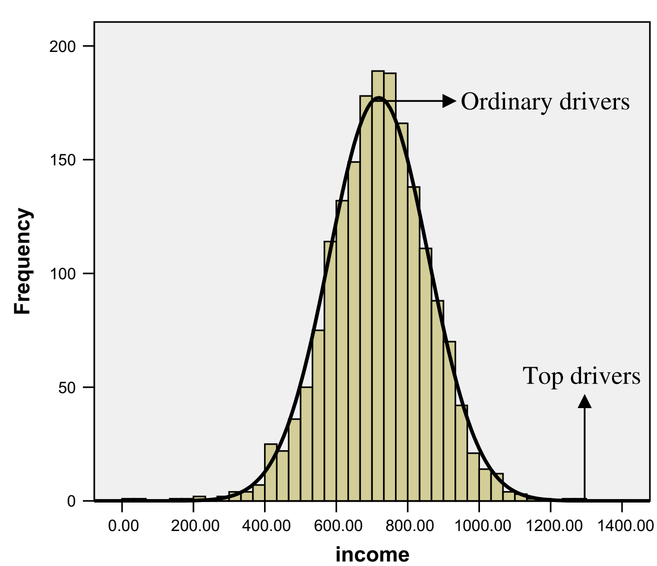

Fig. 1. The daily income of different taxi drivers shows a normal distribution (Unit: Yuan).

[1] Shenzhen Urban Transport Planning Center website, accessed in October 2009: http://www.sutpc.com/szgis/.