Simulation Modelling Practice and Theory 00 (2017) 1–11

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

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

A collaborative calculation on real-time stream in smart cities

Weilong Ding[a,b,*], Shuai Zhang[c], Zhuofeng Zhao[a,b]

[a] Data Engineering Institute, North China University of Technology, Beijing 100144, China
[b] Beijing Key Laboratory on Integration and Analysis of Large-scale Stream Data, Beijing 100144, China
[c] Department of Storage, Dawning Information Industry, Beijing 100194, China

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Collaborative calculation
Real-time data
Historical data
Smart cities
Recognized data of vehicles

ABSTRACT

In the smart cities, the *travel-time* is a typical business calculation to monitor and control the traffic congestions. But it still faces challenges on real-time stream due to the limitation of latency and accuracy. In this paper, we propose a collaborative approach for travel-time calculation on stream of recognized data of vehicles. Compared with other types of sensory data in urban roads, the recognized data of vehicles has wider coverage, finer interval and more exact locality. Our approach continuously achieves both factual and predictive values, and consists of two-step spatio-temporal parallelism on real-time data and Bayes prior rules mining on historical data. It can be analyzed theoretically for its low latency with high accuracy, and has been implemented on Apache Storm correlated with Hadoop MapReduce. Through exhaustive experiments on simulated and real data, our approach holds millisecond-level latencies steadily on high speed stream with nearly linear scalability, and keeps the accuracy above 80% for prediction.

1. Introduction

In transportation domain, there are multiple business calculations in urban road network. The *travel-time* is one of the typical cases [1], which means the average duration of all the vehicles gone through given road segment at given moment. It has been widely used to evaluate current conditions like hot spots or congestion, and provides rational decision supports for road network planning [2] or public travel services [3]. Therefore, the research of travel-time is always the hot areas and has been essential to build intelligent transportation system (ITS) in smart cities.

The principle to calculate travel-time is to find the factual or predictive values at given intervals for all the road segments. The traditional approaches are based on data sample of special floating vehicles [4] or road sensors. In practice, common floating vehicle data could be periodical GPS location of given taxies [5], and road sensors would be inductive loop detectors or devices in SCATS system. However, as special vehicles, the floating ones have limited representativeness, and their GPS quality is relatively low due to unreliable device precision [6]. Moreover, as embedded modules at the build-time of roads, the road sensors are hard to repair and extend, and their data is only for dedicated usage like traffic-volume or travel-time. Nowadays in the large and medium-sized cities, the *recognized data of vehicles* from camera sensors have been adopted to build ITS and public services. A tuple as the unit of recognized data contains multiple attributes including recognized vehicle-plate, passed time, vehicle type, lane number, camera id, and captured photos. Compared with other sensory data in urban roads, the recognized data of vehicles has wider coverage due to the thousands of camera on roads, finer intervals due to the continuous 7×24 functionality, and more exact locality due to the communication among

* Corresponding author.
E-mail addresses: dingweilong@ncut.edu.cn (W. Ding), zhsh1222@126.com (S. Zhang), edzhao@ncut.edu.cn (Z. Zhao).

http://dx.doi.org/10.1016/j.simpat.2017.01.002
1569-190X/

fixed points [7]. On those recognized data, different business calculations could be implemented contemporarily, and the travel-time of Big Data solution is urgently required in academic and industry. In transportation domain, there are two common calculation patterns: the online calculation on the live real-time data and the offline calculation on the accumulative historical data. The former one can reflect real environment more naturally.

However, it still has challenges for the travel-time calculation on recognized data. First, low latency calculation is a problem. In practical condition, the recognized data as stream would be accumulated fast and concurrently from massive sensors at second-level intervals. However, the traditional parallel approaches have to be calculated after the data has been stored, which have the intrinsic shortage due to the physical limitation of disk I/O. Second, accurate calculation is another problem. The uncertain real condition, the exact result cannot be always achieved. For example, the travel-time value may be zero at given intervals or on given road segments due to some data loss or lack. However, the aggregation on Big Streaming Data reflects the continuous fact [8], and this meaningless "zero" should be smoothed by reasonable values in practice.

In this paper, we take travel-time as an example to propose a collaborative calculation on stream of recognized data. Here, both online and offline calculations are included: on live real-time data two-step parallelism pushes factual and predictive values continuously; meanwhile on massive historical data Bayes prior rules are mined periodically to guide the prediction. The contributions of our approach are listed as follows. (1) Two-step spatio-temporal parallelism reduces the latency to millisecond-level utilizing the inherent feature of recognized data; (2) the prediction can soundly improve the accuracy to above 80%; (3) the approach has been implemented for practical use, which has already yielded the business benefits.

This paper is organized as follows. Section 2 shows the scenario and related works. Section 3 elaborates the collaborative approach. Section 4 demonstrates the experiments quantitatively. Section 5 summarizes the conclusion.

2. Background

2.1. Scenario and problem definition

Our work comes from the Grid Transportation project in *Shenzhen*, a big Chinese city. In this city, thousands of cameras are deployed on trunk roads to recognize the passing vehicles for traffic analysis. One tuple would be recognized in about a second by a camera and immediately sent to the backend system. Here, a tuple includes recognized information like vehicle-plate, vehicle type, capture time, lane number, camera number, two photos (front and back view) and 22 other attributes. Fifty billion tuples about 2 TB would be accumulated 1 day, and traditionally 7 h would be expended to scan once. However, it is required that any tuple should respond in no longer than 3 s since its arrival. Accordingly, we developed backend system in Cloud to execute ten more online business calculations, such as fake-plate detection, copy-plate detection, black-list detection, accompanied-vehicle analysis and travel-time. Here, fake-plate detection would find the vehicles whose plates are not legal in each second; copy-plate detection would locate the suspected vehicles with the same plate being too far away theoretically in recent 1 h. We focus on travel-time, a typical aggregation calculation, and have published previous work [9]. Different with those works on massive historical data, the approach in this paper reconstructs the same business on real-time data to provide a low latency solution. Based on the definitions in that work [9], here we extend the Definition 1–3, and append other definitions.

Definition 1. Monitoring point, road segment and road network. A monitoring point is the location where a sensor is deployed to capture recognized data of vehicles. The road segment is the directed path between two adjacent monitoring points: $s_i = \langle l_u, l_d \rangle$. Here, $l_u(l_d)$ is the upstream (downstream) monitoring point. The road network is the network topology composed by road segments: $R = (L, S)$. Here, L is the set of motoring points, $\forall l_u, l_d \in L$, and S is the set of road segments, $\forall s_i \in S$.

Definition 2. Recognized data of vehicles. The recognized data of vehicles is the data captured at monitoring points of the road network. Its basic unit can be abstracted as tuple $u = (v, l, t)$. Here, v is the vehicle-plate of a vehicle, l is the motoring point, and t is the timestamp when v passing l.

Definition 3. Travel-time of single vehicle, travel-time of road segment, and travel-time calculation. The travel-time of single vehicle v_i is the duration that v_i passed through the given segment s_j: $tra_{v_i}^{s_j} = u_2 \bullet t - u_1 \bullet t$, $\exists\, u_2, u_1$, $u_2 \bullet v = u_1 \bullet v = v_i$, and $s_j = \langle u_1 \bullet l, u_2 \bullet l \rangle$. The travel-time of road segment s_j is the mean value of travel-times of all single vehicles passed through s_j at a given time interval δ: $tra^{s_j} = \mathrm{AVG}(tra_{v_i}^{s_j})$. Here, $\forall u, u.t \in \delta$. In practical scenario, δ is always defined as recent 5 min, 15 min or 60 min. Moreover, the "mean value" can be defined in multiple ways under different conditions, and we use the median [10] in this paper. If no emphasis, the travel-time is the one of road segment by default. Accordingly, the travel-time calculation is the procedure to continuously achieve the travel-time of each road segments at every interval δ.

As the discussion in Section 1, the continuous travel-time is meaningful only on a set of travel-time of single vehicles. However, the real conditions make it possible that some roads had few data at late night or early morning. In such cases, their travel-time cannot be deduced by mean value, and the *zero-value* phenomenon appears. In fact, as the continuous value, travel-time reflects the trends in the recent past, and its zero-value is abnormal. Therefore, we adopt the previous predictive value as the substitutes to smooth the travel-time.