Transportation Research Part C 23 (2012) 56–67

Contents lists available at SciVerse ScienceDirect

Transportation Research Part C

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

Sampling methods for summarizing unordered vehicle-to-vehicle data streams

Jiadong Zhang [a,*], Jin Xu [b], Stephen Shaoyi Liao [a,b]

[a] Department of Information Systems, College of Business, City University of Hong Kong, Kowloon, Hong Kong
[b] Department of Electronic Commerce and Information Management, School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, China

ARTICLE INFO

Article history:
Received 30 November 2010
Received in revised form 30 October 2011
Accepted 18 November 2011

Keywords:
Vehicle-to-vehicle environment
Unordered data stream
Summarization
Biased sampling
Traffic state estimation

ABSTRACT

In the vehicle-to-vehicle (V2V) communication environment, vehicles interchange traffic data with each other. Because of the unbounded size of traffic data streams, sampling is used for summarization of traffic data, instead of storing the original data directly, for estimating traffic characteristics such as speed in the next step. All existing sampling methods assume that data arrivals are in the increasing timestamp order. However, this assumption may not be true in the V2V environment due to multiple data sources, transmission delays and different ways of dissemination. This disordered issue is explored in two ways in this paper. First, the traditional sampling methods for ordered streams are extended to be compatible with the disorder, especially the Unordered Extension of Exponentially Biased Reservoir Sampling (UEEBRS). Second, we propose a novel method, called the polynomially biased reservoir sampling (PBRS), to summarize unordered traffic data streams. Two measurements, the relative bias of speed and the cover rate of information obtained from the constructed summarizations, are used to assess performance of the extended methods and the novel way of comparing them with the classical methods. Preliminary simulation results show the proposed methods (UEEBRS and PBRS) reduce the relative bias of speed by about 10% with respect to the best reported result, while their cover rates of information are comparable at least to the others and are sufficiently high to support real-world applications.

1. Introduction

1.1. Background

Traffic congestion has become a growing and intractable issue throughout the world. Traffic congestion costs a massive amount of time, fuel and money (Schrank and Lomax, 2009). In order to solve this problem, the intuitional approach has been to expand roads capacity but usually this is impractical and undesirable because of budget and space constraints. Therefore, attention is being paid to efficient and intelligent utilization of the existing transportation infrastructure. Intelligent transportation systems (ITSs) are intended to provide information to transport infrastructure and vehicles to increase the efficiency of the existing systems, which help reduce commuting time and fuel consumption.

Recently, with the advancement of wireless short-range communication technologies and GPS becoming common practice in vehicle applications, interest in ITS based on vehicle-to-vehicle (V2V) communication or inter-vehicle communication (IVC) (Wischhof et al., 2005; Sen and Matolak, 2008), which is different from the traditional vehicle-to-infrastructure (V2I)

* Corresponding author. Tel.: +852 6737 3684; fax: +852 2788 8694.
E-mail addresses: jiadongzh@gmail.com (J. Zhang), xj_james@163.com (J. Xu), issliao@cityu.edu.hk (S.S. Liao).

0968-090X/$ - see front matter
doi:10.1016/j.trc.2011.11.009

structure, has been growing. V2V-based ITS are implemented in a distributed (decentralized) way, where information can be interchanged through direct communication among vehicles. They have many obvious advantages like less delay and more specific traffic data than conventionally centralized ITS (V2I). Traffic data are essential to data-driven applications such as estimation of traffic states like travel time and traffic flow volume. In particular, smaller delays in data availability are very important for applications that involve little reaction time, such as collision warning (Mitropoulos et al., 2010).

1.2. Research problem

In general, a typical ITS architecture consists of two parts: V2I and V2V. However, this work focuses on the V2V part only, where traffic data are shared among vehicles through periodically broadcast messages using the existing V2V communication technologies (Wischhof et al., 2005; Sen and Matolak, 2008). There are two features: first, it does not require any extensive infrastructure and management centers to collect and disseminate traffic data. Hence no substantial public investment is required for sensing, communication and possibly computing equipment. Secondly, this study is targeted at offering individual vehicles with elegant abilities to pre-process collected data, which make them have potential for further estimating traffic states precisely. Thus it does not matter where the data come from, i.e. other vehicles or infrastructures.

In the V2V environment, vehicles mounted GPS and wireless communication devices can receive signals from GPS satellites to generate traffic data periodically, such as the current location and speed, and then can exchange the traffic data with each other in short-range areas. Therefore, each vehicle can get information of traffic situation around it by analyzing the received traffic data (Wischhof et al., 2005; Sen and Matolak, 2008; Delot et al., 2010). The process of exchanging data is depicted in Fig. 1.

As presented in Fig. 1, for each vehicle, traffic data can be either generated by its own GPS device or it can be obtained from other vehicles. The communicated data can be either a vehicle's own GPS data or data obtained from other nearby vehicles. A GPS data record can contain many attributes, but in this paper, we concentrate on only three attributes: timestamp, location and speed. Thus, the attributes of timestamp and location ensure consistency of the exchanged data. That is, if two records have the same timestamp and location, both are viewed as identical and are processed only once.

As discussed above, data from GPS and other vehicles can be considered as data streams since they possess the general characteristics of data streams: massive volume of data, and temporal correlations. Firstly, as time goes on, volume of the acquired data is potentially unbounded (Zhang et al., 2010). Secondly, the data (timestamp, location, speed, etc.) contain the time attribute, which indicates the precise time when the GPS data are generated, i.e. the point of time for which the data represent the traffic situation. Actually, data streams are essentially highly dynamic time series (Ogras and Ferhatosmanoglu, 2006).

The volume of such data is so large that it may be impossible to store the data (Aggarwal, 2007; Zhang et al., 2010). In particular, memory of vehicle-mounted devices with GPS and other communication technologies is quite limited. Thus, the incoming data need to be processed in real time. After that, the data are discarded. Further, even though the data can be stored, the volume of the incoming data may be so large that it would be impossible to process any particular record more than once because of the limited computational power, especially of vehicle-mounted devices (Aggarwal, 2007).

Fortunately, summarizations can be constructed and updated from data streams as they are received. Then the summarizations, instead of the original data, are used for estimations of traffic states in a variety of terms, such as traffic flow volume and travel time. In our previous study (Zhang et al., 2010), we implemented the framework to summarize V2V data streams (Fig. 2). First, GPS and other wireless communication devices fitted in vehicles collect and communicate the data, which are then queued into typical data streams, in accordance with their arrival time. Secondly, in order to address this large volume of data in real time, three sampling methods, including sliding window sampling (SWS), reservoir sampling (RS) and exponentially biased reservoir sampling (EBRS) in data stream theory, are used to provide the users with summarizations for further analysis.

The sliding window model is motivated by the idea that users are more concerned with analysis of the most recent data (Datar and Motwani, 2007). From another perspective, reservoir sampling views all data as equally useful. That is, all records are included in the sample with the same probability (Vitter, 1985). However, the two sampling methods stand for two ex-

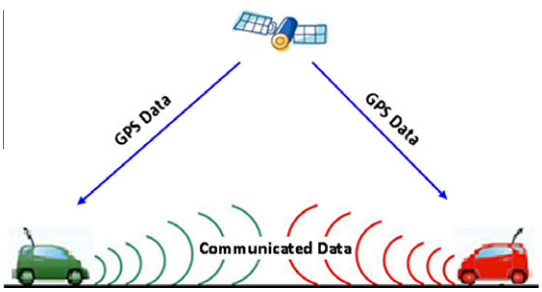

Fig. 1. Data exchange among vehicles in V2V environment.