Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Monte Carlo simulation-based traffic speed forecasting using historical big data

Seungwoo Jeon, Bonghee Hong*

Department of Electrical and Computer Engineering, Pusan National University, Busan, South Korea

HIGHLIGHTS

- We propose a three-step filtering algorithm to find and remove data outliers.
- We use various prediction accuracy measures to verify our statistical model.
- Our statistical model has a high prediction accuracy for each road.
- We construct a big data processing framework to handle the overall process.

ARTICLE INFO

Article history:
Received 29 May 2015
Received in revised form
24 October 2015
Accepted 21 November 2015
Available online xxxx

Keywords:
Big historical traffic data
Changepoint analysis
Correlation analysis
Monte Carlo simulation
Accuracy measures

ABSTRACT

Because the traffic patterns on roads vary according to the roads' specific spatio-temporal behavior, if we would like to forecast the traffic speed by day of the week, it is necessary to determine an optimal set of the highly related historical patterns to achieve high prediction accuracy. The goal of our paper is to suggest a new statistical modeling method that finds the best historical dataset according to various analyses for each link and provides a more accurate prediction of traffic flow by day of the week. First, we suggest a three-step filtering algorithm based on changepoint analysis, correlation analysis, and Monte Carlo simulation to simultaneously find and remove historical data outliers. Second, we determine the optimal historical data range by using decision factors such as the Mean Squared Error (MSE) and Akaike Information Criterion. Moreover, to verify our statistical model, we use various prediction accuracy measures such as Mean Absolute Percentage Error (MAPE), R-squared value, and Root MSE (RMSE). Finally, we construct a big data processing framework to handle the overall prediction process and calculate large amounts of traffic data. The forecasting results show that the proposed model can achieve a high prediction accuracy for each road by using three measures: less than 20% for MAPE, more than 80% for R-squared value, and less than 1 on average for RMSE.

1. Introduction

We live among a variety of big data sources such as data from GPS signals, closed-circuit television (CCTV), traffic loop detectors, climate sensors, credit card payment information, and social network services [1,2]. In particular, we can know the location of currently congested roads and find shortest routes in real-time by using real-time traffic data. These data have been provided by Intelligent Transportation System (ITS) sensor devices like Vehicle Detection Systems and Dedicated Short-Range Communications (DSRC) over the past few decades [3–6]. Recently, because of the growing number of vehicles and indiscriminately congested roads, most people would like to know the traffic conditions of the next day or day of the week in advance. Currently, we should be able to estimate the travel time of routes composed of links (e.g., separated at an intersection where traffic flow changes) if we predict traffic speeds for all times and links connecting from a start node to an end node.

However, the more difficult job of traffic flow prediction is to predict the traffic flow of the next day of the week in addition to the continuous prediction of on-going traffic flow. The prediction of on-going traffic flow is an easier job than non-continuous prediction. For example, suppose that we would like to predict the flow within one hour by using real-time data. Because the traffic flow does not unexpectedly change, it is easy to predict the flow by using one-hour old or current data. On the other hand, when we would like to predict the rush hour flow on the next day or next Monday, it is more difficult to predict the non-continuous traffic flow from

* Corresponding author.
E-mail addresses: i2825t@pusan.ac.kr (S. Jeon), bhhong@pusan.ac.kr (B. Hong).

http://dx.doi.org/10.1016/j.future.2015.11.022
0167-739X/

historical traffic information because of the many variables affecting the traffic flow, including traffic congestion that could happen over the next couple of days. Namely, the prediction of future traffic flow over the next day of the week depends upon an irregular pattern of traffic flow.

To clarify the problem of prediction, we refer to the prediction of on-going continuous traffic flow as continuous prediction (CP) because it uses on-going continuous traffic data within the same day. On the other hand, because the historical pattern of traffic flow is used for predicting a particular day of the week, we refer to the non-continuous prediction over the upcoming day of the week as non-continuous prediction (NCP). Recently, some ITSs have begun storing historical data and attempting to predict the travel time of congested roads by using this data [7,8]. However, it is difficult to choose an optimal set of historical data to use as input in non-continuous prediction since the prediction accuracy of a day's traffic can vary depending on the choice of the data. In general, the historical data consists of two patterns: normal patterns and abnormal patterns, which are completely different patterns caused by specific events such as festivals or accidents during specific time intervals. Moreover, because of road construction over a certain period of time, even traffic flow with a normal pattern could be completely changed afterward. This is called a big change pattern. These abnormal and big change patterns are outliers that can considerably affect the prediction accuracy.

To solve the historical data selection issue, the historical data should be classified with other data that have normal patterns while outliers are simultaneously removed. In this paper, we propose a three-step filtering algorithm that first determines large changes in the patterns and excludes data before these large changes. It then removes historical data with low correlation coefficients. The final step randomly combines the remaining data by using Monte Carlo simulation to determine the best input combination. Furthermore, we finalize the data selection by determining the optimal historical data range, using the decision factors of each method. For example, suppose that we have one hundred historical data (2014.01.01 to 2014.04.10) in a link that sharply changes on 2014.02.16 because of road construction and for which the flow between weekends and weekdays is completely different. In the first step, we exclude fifty data (2014.01.01 to 2014.02.16) because of the big changes in the patterns. To predict the flow of 2014.04.06, thirty-six weekday data are excluded in the remainder of the data according to correlation analysis. Lastly, thirteen data as final input data are selected by the simulation and decision factors.

In addition, we suggest a two-step verification to select the optimal time series forecasting methods. In the first step, the Mean Absolute Percentage Error (MAPE) of each method generated by simulation is compared and the predicted data combination with the smallest MAPE is selected. We determine the final optimal time series forecasting method by using three measures of the difference between the predicted data and forecasting day. In this paper, because of the continuously increasing historical traffic data on all roads, we focus on the "volume" aspect of big data. It is possible to predict traffic speed by selecting the optimal historical dataset from the big data, using distributed parallel processing.

The main contributions of this paper can be summarized as follows.

- We propose a new statistical model for generating prediction data with high accuracy, removing input data outliers through correlation analysis and Monte Carlo simulation, applying a time-series forecasting method to each link and day, and generating the best prediction data for each time duration by using the Mean Squared Error (MSE) and Akaike Information Criterion (AIC). Finally, we verify our modeling by using the cross-validation of three measures: MAPE, *R*-squared, and Root MSE (RMSE). These measure prediction accuracy, that is, they measure how well the data are predicted.
- We construct a forecasting system based on big data open source tools. This system consists of Hadoop, RHive, and Hive for data processing and R as the statistical analysis package. It performs all analysis from historical data insertion to verification of the predicted data. Furthermore, because it is necessary to reduce loads that frequently call MapReduce jobs to use raw data stored in Hadoop during various analyses, we store the raw data in an R data file to increase the processing speed of analyses and calculations.

The remainder of this paper is organized as follows. Section 2 presents the background to understanding traffic data and analyzes the characteristics of that data. Section 3 defines the problems and limitations of research on data and methods. In addition, Section 4 and 5 describe our statistical model and system architecture in detail via an example. Section 6 presents our experiments that use several scenarios. In Section 7, we review various existing research for time series forecasting methods. Section 8 presents a summary of our approach and contributions. Finally, Section 9 concludes the paper.

2. Traffic data basics

In this section, we first introduce the standard concepts of nodes and links in ITS, and then describe the traffic data collected from ITS. Finally, we present various traffic patterns.

2.1. Traffic data specification

Traffic data indicate the average speed of vehicles in a standard link. A link is defined on a line that connects a standard start node to a standard end node where traffic flow changes, such as at an intersection. Table 1 shows an example of real traffic data, which is composed of link ID, time, number of cars, and average speed. The data are produced by combining and smoothing several raw traffic data observations such as those of buses, taxis, and DSRCs. The link ID field is the link's unique identification number, and time is recorded in the "yyyymmddhhmmss" format in five-minute intervals. The third field is the number of vehicles passing through the link. The average speed is the final field, in which speed is computed using the average speed of all lanes on a specified road (link).

2.2. Historical traffic data as big data

We were able to secure a big dataset of historical traffic data, as detailed in Table 2. The real data was collected by the Busan city government from Nov. 2013 to Jan. 2015. The area of collection covers all roads in Busan, Korea. The scale of the data collected is 3–4 GB per month; in total, the dataset is 55 GB so far. Considering only size, 55 GB may not be enough to classify this dataset as big data. However, since 3–4 million observations are collected in one day, eventually, over one billion observations will be collected.

2.3. Characteristics of traffic data

The traffic data for each link have various patterns according to time, space, and events, as shown in Fig. 1. While all traffic patterns over one week on a link are different except during the night time in Fig. 1(a), (b) shows almost similar patterns between two days separated by a one week interval. Fig. 1(c) describes different patterns that depend on space; whereas the speeds of a downtown street are overall low because of many vehicles, the speeds of an uptown street are high, regardless of time. Finally, Fig. 1(d) presents patterns between normal and abnormal days such as festival events. Both patterns are similar until the afternoon, but for the festival day the speed is temporarily low from evening

Please cite this article in press as: S. Jeon, B. Hong, Monte Carlo simulation-based traffic speed forecasting using historical big data, Future Generation Computer Systems (2015), http://dx.doi.org/10.1016/j.future.2015.11.022