IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 30, 2018, accepted September 27, 2018, date of publication October 4, 2018, date of current version October 29, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2873569

Congestion Prediction With Big Data for Real-Time Highway Traffic

FAN-HSUN TSENG[1], (Member, IEEE), JEN-HAO HSUEH[2], CHIA-WEI TSENG[2], (Student Member, IEEE), YAO-TSUNG YANG[2], (Student Member, IEEE), Han-Chieh Chao[3,4], (Senior Member, IEEE), AND LI-DER CHOU[2], (Member, IEEE)

[1]Department of Technology Application and Human Resource Development, National Taiwan Normal University, Taipei 106, Taiwan
[2]Department of Computer Science and Information Engineering, National Central University, Taoyuan 320, Taiwan
[3]Department of Electrical Engineering, National Dong Hwa University, Hualien 974, Taiwan
[4]Department of Computer Science and Information Engineering and the Department of Electronic Engineering, National Ilan University, I-Lan 260, Taiwan

Corresponding author: Li-Der Chou (cld@csie.ncu.edu.tw)

This work was supported in part by the Young Scholar Fellowship Program by the Ministry of Science and Technology (MOST), Taiwan, under Grant MOST107-2636-E-003-001, and in part by the MOST under Grant MOST102-2221-E-008-039-MY3, Grant MOST103-2221-E-008-090-MY3, Grant MOST104-2221-E-008-039-MY3, Grant MOST105-2221-E-008-071-MY3, and Grant MOST107-2221-E-259-005-MY3.

ABSTRACT By collecting and analyzing a vast quantity and different categories of information, traffic flow and road congestion can be predicted and avoided in intelligent transportation system. However, how to tackle with these big data is vital but challenging. Most of the existing literatures utilized batch method to process a bunch of road data that cannot achieve real-time traffic prediction. In this paper, we use the spouts and bolts in Apache Storm to implement a real-time traffic prediction model by analyzing enormous streaming data, such as road density, traffic events, and rainfall volume. The proposed SVM-based real-time highway traffic congestion prediction (SRHTCP) model collects the road data from the Taiwan Area National Freeway Bureau, the traffic events reported by road users from the Police Broadcasting Service in Taiwan, and the weather data from the Central Weather Bureau in Taiwan. We use fuzzy theory to evaluate the traffic level of road section in real time with considering road speed, road density, road traffic volume, and the rainfall of road sections. In addition, the SRHTCP model predicts the road speed of next time period by exploring streaming traffic and weather data. Results showed that the proposed SRHTCP model improves 25.6% prediction accuracy than the prediction method based on weighted exponential moving average method under the measurement of mean absolute relative error.

INDEX TERMS Big data, fuzzy theory, intelligent transportation system, real-time streaming data, support vector machine.

I. INTRODUCTION

According to the study of International Data Corporation, the usage of digital data worldwide is about 1.8 zettabytes in 2011. The study further predicts that data amount will be 44 times more than nowadays in 2020, about 35.2 zettabytes. These digital data are generated by a variety of different ways. The online auction company eBay achieves online transactions in millions every day. There are more than 88 million users and more than millions of merchandise queries so that eBay's database increases more than 50 terabytes data every day. To analyze user behavior, the online system of eBay deals with more than 50 petabytes data and executes more than 5 thousands items of business analysis per day. The enormous amount of data information is regarded as big data [1].

The well-known technology research Gartner points out that big data should be provided with high capacity, high growth capacity and high variability of characteristics. In 2001, Doug [2] pointed out that there are three data growth directions, i.e., volume (data size), velocity (data transfer speed) and variety (diversity of information). Big data feature. In [3], Chandarana and Vijayalakshmi pointed out that the characteristics of 5V of big data.

- *Volume*: A great quantity of data has been created quickly since everyone has one or more than one mobile devices. The data size is large and growing quickly reached terabytes even petabytes.
- *Velocity*: The data will be generated rapidly. In big data, the data generation and deletion, data flow, data change, and data processing are fast beyond our imagination.
- *Variety*: Based on data structure, big data can be classified into two categories, i.e., structured and unstructured data. There are various diversities of data types and forms, e.g., data from different mobile phones, different social media data like Facebook or Twitter, data from sensor networks, vehicle-to-vehicle and device-to-device networks.
- *Veracity*: The uncertainty and reliability of big data is a big issue. The unstructured data usually has problems in imprecision of data. For instance, text usually has more than one meaning. Therefore, big data should be analyzed preciously.
- *Value*: The value of big data is a vital issue. How to explore the real value of big data by processing enormous and varied data is an important research topic.

The growing speed of automotive industry is as fast as big data. The Organisation Internationale des Constructeurs d'Automobiles (OICA) [4] pointed out that there are 70 million cars produced worldwide every year in recent years. It even reaches 90 million cars in both 2013 and 2014. According to Ministry of Transportation and Communications (MOTC) Republic of China (R.O.C.) [5] statistics, vehicle usage in highways attains to the amount of more than 530 million every year during 2010 to 2013. Since 2014, the National Freeway Bureau in Taiwan launches its new ''Pay as You Go'' toll system. As a result, it can be observed that highways play an important role in Taiwan no matter in cities, towns or rural areas. However, the growing highway usage raises the probability of traffic jam, e.g. periodical congestion sections and sudden traffic accidents. Various literatures investigate traffic jam issues but most of them focus on historical data analysis. On the other hand, existing literatures dichotomize traffic status into traffic jam or not as their prediction results. It is unable to describe driver's feeling of congestion event precisely. Unlike existing literatures, the contribution of this work is as follows.

- The paper proposes an SVM-based Real-time Highway Traffic Congestion Prediction (SRHTCP) model that instantaneously forecasts the car speed of next time period and analyzes traffic jams in highways.
- The traffic analysis and prediction is accomplished by collecting different data formats and sources. To deal with traffic data, more than 150 thousand data in Taiwan Area National Freeway Bureau (TANFB) collected from 3617 vehicle detectors (VDs) along highways in Taiwan, 9.6 thousand real-time weather data from the Central Weather Bureau of Taiwan (CWBT) and social media from the Police Broadcasting Service of Taiwan (PBST) are processed.
- The paper uses Apache Storm [6] to process real-time streaming data, and utilizes fuzzy theory to analyze driver's feeling of traffic jam. The proposed SRHTCP model is superior to other methods in terms of prediction accuracy.

The rest of this paper is organized as follows. Section II discusses and compares related works. In Section III, the designed system based on Apache Storm framework is introduced. Section IV illustrates experiment and prediction results. Finally, Section V concludes this work.

II. RELATED WORKS

Based on the concept of big data, this work collects and analyzes various data types and formats to predict real-time highway traffic. Several technologies are used in the paper, i.e., Apache Storm, traffic theory, fuzzy theory and support vector machine.

A. APACHE STORM

Apache Storm [7] is an open source and distributed real-time computing system. It can easily and quickly process the undivided streaming data and has following features.

- Widely used: Apache Storm is able to process message and update database. Moreover, it can continuously query and report streaming data thus proceed with a great number of requests.
- Scalability: Apache Storm is provided with excellent scalability that enormous machines can be adjusted and configured at the same time. It achieves one million data processing per second by using 10 cluster nodes equipped with Apache Storm.
- Availability: A real-time system must ensure that each data are proceeded successfully. Apache Storm traces every data through its message ID to guarantee data availability.
- Robustness: The goal of Apache Storm is easy management. The administrator easily obtains user experience and monitors machines conveniently.
- Fault tolerance: While a machine is out of order, Apache Storm can reboot the broken machine without influencing other on-line machines. It guarantees that a task can be executed infinitely unless the task is terminated manually.
- Variety of programming languages: Apache Storm is robust and flexible so that developers can program it with multiple different programming languages.

Apache Storm is composed of Nimbus, Supervisor and Zookeeper. Nimbus is the brain of Apache Storm cluster and runs on the master node. It is responsible for sending tasks to other nodes and monitoring operation status of the cluster. Note that there is only one Nimbus in the whole cluster. Supervisor tackles task reception and monitoring. It runs on all working nodes and turns on or off task process based on its received task. Zookeeper plays the communicator role in

Apache Storm cluster. All nodes send heartbeat message to Zookeeper.

The particular design of Apache Storm achieves real-time streaming data processing. The task submission program in Apache Storm is called *Topology*. A topology is composed of several spouts and bolts. The smallest processing unit is called tuple. These three components are introduced as follows.

- Tuple: The smallest unit to compose a stream. In the topology of Apache Storm, all data are transmitted with stream format.
- Spout: An interface between topology and streaming data. When data are settled and mapped, data spout to topology and forwarded to bolt for processing.
- Bolt: A bolt is an element for data processing in topology. Data calculations are conducted in bolts. A bolt forwards data to another bolt after its data processing.

The relation between spouts and bolts in a topology is illustrated with Fig. 1. When a spout receives streaming data, it divides the data into several tuple and forwards to corresponding bolts. Note that different bolts tackle with different data processing then output data or send to next bolt. Therefore, we need to design different topologies to accomplish different data processing.

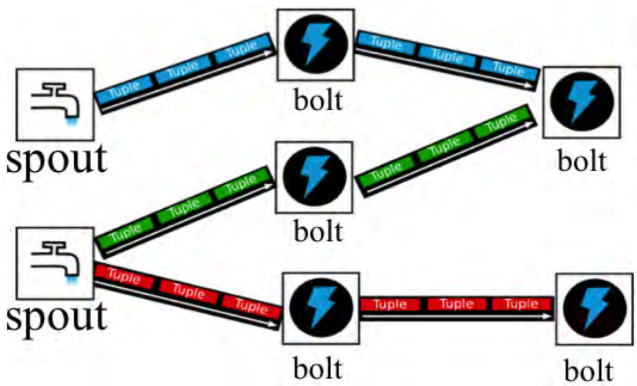


FIGURE 1. The relation between spouts and bolts in topology.

B. OPEN DATA

In Taiwan, government provides almost 30 thousand open data sets in the open data platform [8], e.g., weather, traffic, infrastructure, education, construction, election. In the paper, traffic data, weather data and social media data are integrated and analyzed.

- Traffic open data: The paper retrieved the traffic data of highways in Taiwan from the TANFB [9]. Open data in the platform includes speed of sections, VDs, closed-circuit televisions, changeable message signs and automatic vehicle identification. We use the information provided by VDs in this paper.
- Weather open data: The paper adopted weather data from the CWBT [10]. Open data in the platform includes weather forecasts, observation, earthquake, tsunami, climate and weather. We use rainfall information from weather stations in this paper.
- Social media open data: The paper collects social media data from the PBST [11] so that the prediction tallies with real-time traffic status. Drivers are capable of informing traffic reports and events to PBST. Four kinds of traffic report are recorded in PBST, i.e., traffic barriers, road construction, traffic congestion and other events. In this paper, we use real-time traffic reports in PBST.

C. TRAFFIC THEORY

In traffic theory, traffic stream models are used to represent the relation between volume, speed and density. In [12], Greenshields proposed a parabolic function and named it as Greenshields model. The Greenshields model describes that the higher density and volume result in the lower car speed, and vice versa. It means that the probability of traffic congestion is higher while car density is rising. In [13], Nakayama *et al.* verified the relation between car density and speed with some experiments. The experiment is implemented in a ring road without any obstacle. The experiment is designed to examine that whether car volume influences car speed or leads to traffic accident or not. The results showed that traffic jam happens when there are 22 cars in a 230 meters ring road. Rainfall is an important factor that impacts on traffic flow. In [14], the most congested M42 highway in United Kingdom was investigated for analyzing drivers' behaviors. The researchers found that a sudden slowdown or change lines gives rise to the reason of traffic jam.

D. FUZZY THEORY

In 1965, Zadeh firstly proposed fuzzy theory [15]. The fuzzy theory was proposed to solve uncertainties in the real world by using computers. It uses Fuzzy Control Logic in its mechanism, includes fuzzification, fuzzy database, fuzzy inference and defuzzification. In the fuzzification, input parameters are converted into the membership level of fuzzy sets by the membership function. Then, various "if-then" judgment equations are pre-defined in fuzzy database. Fuzzy inference estimates the membership level of input parameters based on the defined fuzzy database. Lastly, defuzzification step converts inputs into numerical results so that computers are able to determine final result. In the paper, we utilize fuzzy theory to evaluate traffic jam level.

E. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) [16] is a well-known classification technique. It can be applied to solve various research issues, e.g., virtual machines classification in clouds [17], anomaly detection [17], data diagnosis [18] and sensor fault classification [19]. In classification process, the main concept of SVM is to construct an optimal hyper-plane to be the boundary for making decisions. Three characteristics of SVM are as follows. (i) It establishes the maximum boundary so that the coordinates of cluster nodes have the maximum possible distance from the decision boundary. (ii) It uses kernel function to establish a separated linear hyper-plane, thus cluster nodes in higher dimension can be divided into multiple clusters easily. (iii) It adopts a small part of data as training data so that the prediction can be more accurate after training process. In the paper, we utilize SVM to forecast the car speed of highways in Taiwan.

F. COMPARISON OF RELATED WORKS

During past years in Intelligent Transportation System (ITS), various researchers have utilized different methods to monitor traffic events [21], [22] and to predict traffic congestion [23]–[27]. In [21], Milojevic and Rakocevic proposed a vehicle-to-vehicle congestion detection algorithm based on the IEEE 802.11p standard. The proposed algorithm permits vehicles to be self-aware so that vehicles are able to monitor speed and cooperate with each other. In [22], Cheng *et al*. proposed a new automatic incident detection method for urban expressways based on geometric conditions and detector locations. However, these two articles only focus on traffic detection rather the congestion prediction in this paper.

In [23], Ji *et al*. utilized Kalman filter with the Global Positioning System (GPS) location reported by drivers to forecast travel time dynamically. The results showed that the prediction accuracy of improved model is superior to the original method with Kalman filter. In [24], Feng *et al*. also utilized Kalman filter but to predict vehicle's future location. Results showed that the proposed method is superior to a prediction method based on neural network. However, both two methods did not consider weather data and real-time traffic events. In addition, the authors investigated travel time prediction and vehicle's location prediction rather than traffic congestion prediction in this paper. Moreover, the researchers in [23] did not consider highway.

In [25], Wang used Grey prediction to detect traffic incidents. The researcher used actual examples to compare the difference between prediction and reality. Results showed that the proposed method achieves acceptable false-alarm rate to determine whether incident happened. However, the paper is different to our work that traffic incidents are collected from PBST to forecast traffic jams in next time period. In [26], Kuo applied the Kalman filter to Support Vector Machine for achieving travel time prediction. The air pollution index was considered in the prediction model to fit in with real traffic status. However, the author did not consider traffic incidents and real-time weather data. In [27], Li *et al*. proposed a bipolar traffic density awareness routing protocol for vehicle ad hoc networks. The average inter-vehicle space of vehicle networks was predicted, but both traffic events and weather data were unmentioned. In addition, the category and scale of collected data in these three papers is less and much smaller than our work. Moreover, they did not collect real-time traffic and weather data which is achieved in our work. Most of existing literatures used a batch of traffic data to predict car speed, however it cannot achieve instantaneous traffic prediction.

Some researchers have utilized Apache Storm to process streaming data. In [28], Wang *et al*. used Apache Storm their main processing system. The simulation-based result showed that the system can be applied to practical situation. In [29], Chardonnens *et al*. utilized Storm to realize real-time integration and detection on Twitter [30] message keyword statistics and Bitly [31] short Uniform Resource Locator (URL) location statistics. In [32], Bifet and Morales proposed an open source platform called Scalable Advanced Massive Online Analysis (SAMOA). The proposed SAMOA integrates Storm for mining big data streams. In this paper, we utilize the Apache Storm platform to collect traffic data, weather data and social media data instantaneously, and then predict the car speed of freeways in real-time.

III. SYSTEM DESIGN

Based on the concept of big data, the paper proposes a mechanism that uses Apache Storm system to collect and analyze freeway traffic information, weather information and social network information. The proposed SRHTCP model utilizes SVM to forecast the speed of next road section, and instantaneously evaluates freeway condition by using fuzzy theory.

A. SYSTEM ARCHITECTURE AND WORKFLOW

The system architecture of this work is captured in Fig. 2. The proposed cluster architecture is based on the open source distributed real-time computation system *Apache Storm*(or simply Storm throughout this paper). With the master node, it contains Nimbus and user interface, Zookeeper for data exchange. The slave node is composed of supervisor and worker. The proposed congestion prediction method for real-time freeway traffic is implemented in the Storm topology framework. It is composed of six components, i.e., data spout, traffic data bolt, weather data bolt, social media data bolt, data mapping bolt and congestion prediction bolt. The traffic data bolt calculates vehicle density on roads then sends

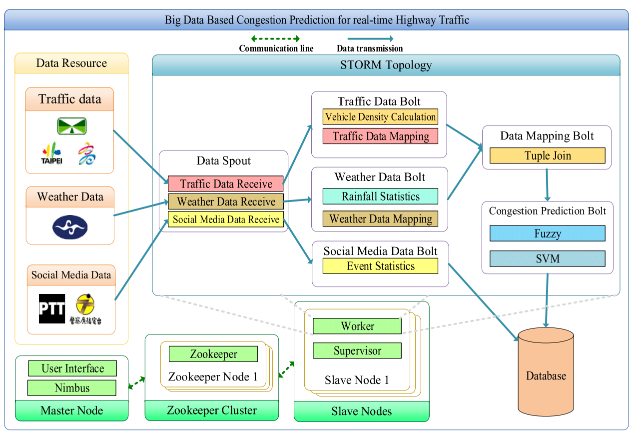


FIGURE 2. System architecture.

the calculation results to the traffic data mapping module. The weather data bolt gathers rainfall statistics then examines and sends to weather data mapping module. The social media data bolt collects four types of road events then sends to database.

The computing cluster of Storm is composed of master node, slave node and Zookeeper. The master node is regarded as the brain of Apache Storm that executes the Storm Nimbus procedure. It submits and distributes all tasks by the Nimbus, and provides administrator with user interface to manage Storm system. The slave node executes Strom Supervisor procedure that distributes tasks at any time. The worker module is launched once the slave node receives task. The Zookeeper plays the communicator role in Storm cluster and records the system status.

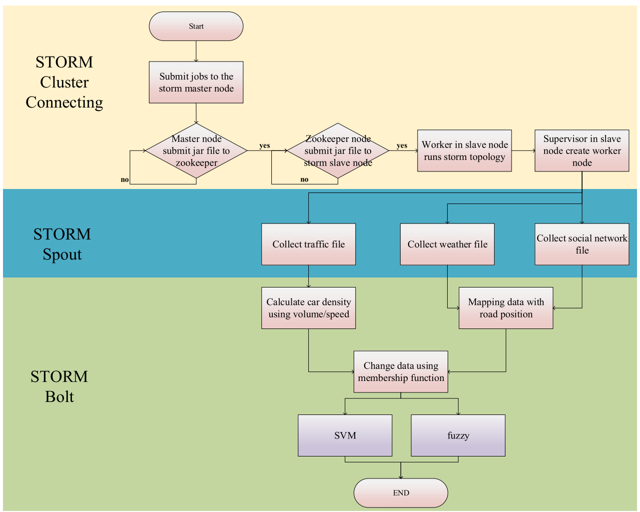


FIGURE 3. **System workflow.**

The system workflow is shown as Fig. 3. The constructed Storm cluster is responsible for job execution by creating master and slave node. In slave node, worker nodes are created to run Storm topology. After that, the Storm spout collects traffic, weather and social network data from the TANFB, CWBT and PBST. In the Storm bolt layer, Fuzzy theory is utilized to analyze real-time traffic congestion level, and SVM is utilized to forecast car speed of lane in next time period.

B. DATA SPOUT AND BOLTS

The Apache Hadoop applies batch method to separately process enormous data, however it cannot immediately process data. As a result, Apache Storm was proposed to deal with real-time streaming data through the spout and bolt. The designed data spout and bolts are explained as follows.

1) DATA SPOUT

The data spout for data collection is composed of traffic data receive module, weather data receive module and social media data receive module. The traffic data receive module is responsible to retrieve data from Taiwan Area National Freeway Bureau to Storm system. Firstly, the module calls the open() function in spout to set the traffic data path for loading the XML file about traffic status. Then, the nextTuple() function reads data from the XML file one by one. A piece of data includes device ID, lane number, vehicle speed in lane and the number of vehicles with different types. Note that four types of vehicles are used in the paper, i.e., motorcycle, compact, van and trailer.

The weather data receive module collects weather data from CWBT to Storm system. The module calls the open() function in spout to retrieve the weather of a location from the XML file about weather information, then reads data by using nextTuple() function from the XML file one by one. Note that the module retrieves the rainfall value of selected location in every ten minutes.

The social media data receive module collects media data from social networks for reporting traffic conditions. The retrieve method is same with the other two modules but from the JSON file. The collected media data includes the information of event such as location, freeway number, direction and type. For instance, a piece of data reported from social media data is recorded as freeway number 1, 235 km at southbound, and car accident.

2) TRAFFIC DATA BOLT

The traffic data bolt for data processing contains vehicle density calculation module and vehicle speed calculation module. The vehicle density calculation module receives tuple data from the data spout by using execute() function. The traffic volume is formulated as

$$q = ku, \tag{1}$$

where q is traffic volume as well as the number of vehicles per hour, k is road density that represents the average number of vehicles in one kilometer, and u is car speed in kilometer per hour. The obtained road density is attached behind the tuple data of traffic data receive module and sent to traffic data mapping module with new tuple format.

The traffic data mapping module receives the tuple data from vehicle density calculation module. It loops up the value in density field from the tuple data, and retrieves the corresponding latitude and longitude then adds to the tuple data. The updated tuple data is sent to data mining bolt by using the emit() function.

3) WEATHER DATA BOLT

The weather data bolt contains rainfall statistics module and weather data mapping module. The rainfall statistics module receives weather tuple data from data spout by using execute() function. It examines the correctness of the weather tuple data. If the rainfall value is less than zero, the weather station observes none of rain so that the negative value should be revised to zero. The updated rainfall value in weather tuple data is sent to weather data mapping module.

The weather data mapping module receives weather tuple data from the rainfall statistics module and retrieves latitude and longitude value. Then it searches vehicle detectors with 1 km distance from the weather station, and adds vehicle detector ID to the weather tuple data. The updated weather tuple data is transmitted to the data mapping bolt by using the emit() function.

4) OTHER BOLTS

In this subsection three bolts are introduced, i.e., social media data bolt, data mapping bolt and congestion prediction bolt. In social media data bolt, the event statistics module uses execute() function to receive tuple data from data spout for data selection. The tuple data includes four types of road events, i.e., traffic congestion, traffic barrier, road construction and other events. The event statistics module uses emit() function to transfer road type reports to database.

In data mapping bolt, the tuple join module receives vehicle tuple and weather tuple data from the traffic data bolt and weather data bolt respectively. It integrates vehicle detector tuple and weather tuple with the same vehicle detector IDs. The integrated tuple data records vehicle detector's status, lane ID, average speed of lane in kilometer per hour, vehicle type, and the volume of lane. There are three different statuses of a vehicle detector, i.e., functional, time-out and malfunction. The lane volume represents the number of vehicles in a lane in one minute.

The fuzzy module and SVM module are in congestion prediction bolt. The fuzzy module converts input data into different membership functions of fuzzy set. The fuzzy inference is executed and completed based on fuzzy database. The output data is the membership between traffic congestion level and other parameters. Last, the fuzzy module quantifies output results through defuzzification and acquires the real-time traffic congestion level. On the other hand, the SVM module predicts the speed of next time period based on the received tuple data from tuple join module. The historical data includes real-time speed, the speed of 5 and 10 minutes before.

C. DATA PROCESSING AND PARAMETER SELECTION

With the vehicle data, Taiwan Area National Freeway Bureau provides data from vehicle detectors every 5 minutes. The used parameters and the flowchart of vehicle data processing are listed in Table 1 and captured in Fig. 4. In the beginning, system tries to access data in Taiwan Area National Freeway Bureau. If the system cannot retrieve data in time, it will wait for a time period to access in next time period. After obtaining the vehicle information by using XML parser to read the data, the XML parser segments data into several data sets based on vehicle detector. The parameter VD_status_x represents the status of vehicle detector x. If $VD\_status_x = 1$, the connection between vehicle detector x and Taiwan Area National Freeway Bureau is disconnected. If $VD\_status_x = 2$, the vehicle detector is malfunction. The output data are negative in both statuses. In order not to affect the fuzzy and SVM modules, the parameters $Speed_{x,lane}$ and $Density_{x,lane}$ of this vehicle detector are set to 1. If the parameter $VD\_status_x = 0$, the vehicle detector is functional so that the obtained $Density_{x,lane}$ and $Speed_{x,lane}$ are sent to fuzzy module and SVM module respectively. The parameter $Density_{x,lane}$ is calculated as

$$Density_{x,lane} = \frac{Volume_{x,lane,car\_all}}{Speed_{x,lane}}, \tag{2}$$

where $volume_{x,lane,car\_all}$ is the obtained car volume of a line from vehicle detector x.

With the weather data, the Central Weather Bureau in Taiwan provides information from rainfall detector every ten minutes. The used parameters and the flowchart of weather data processing are listed in Table 2 and captured in Fig. 5. First, the system tries to access data from Central Weather Bureau, and waits for next time period if it fails to access the data. After obtaining the weather data by using XML parser, the XML parser segments data into several data sets based on rainfall detector. If the parameter $Value\_10_x < 0$, the connection between rainfall detector x and Central Weather Bureau is incorrect. To avoid influencing fuzzy module, the rainfall

TABLE 1. Parameter of vehicle data.

Parameter	Description
X	Total number of vehicle detector
VD_URL	Traffic file's URL
ID_x	Vehicle detector ID
VD_status_x	The status of vehicle detector x
$lane$	Total number of lanes
$Speed_{x,lane}$	Average speed for each lanes and ID_x
$Volume_{x,lane}$	Car volume of each lane for ID_x
$Density_{x,lane}$	Density of each lane for ID_x
$Data\_time$	Data collected time

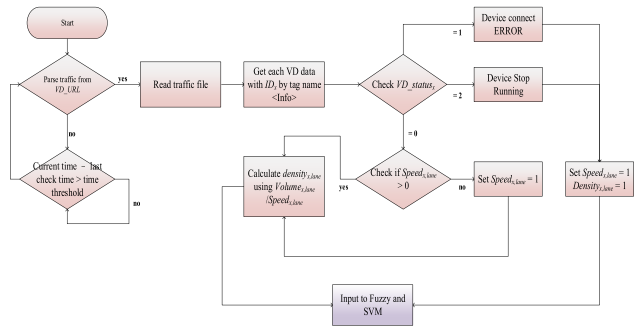


FIGURE 4. Flowchart of vehicle data processing.

TABLE 2. Parameter of weather data.

Parameter	Description
ST_URL	Weather file's URL.
X	Total number of rainfall detector
ID_x	Rainfall detector's ID
ST_lat_x	Latitude of rainfall detector x
ST_lon_x	Longitude of rainfall detector x
$Value\_10_x$	Rainfall value in 10 minute at rainfall detector x
$Data\_time$	Data collected time

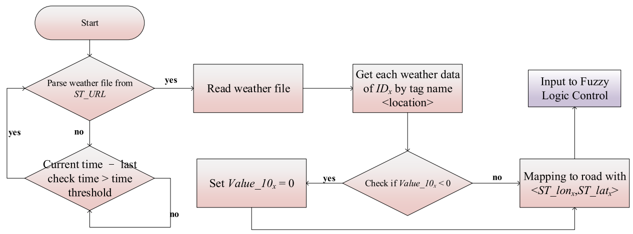


FIGURE 5. Flowchart of weather data processing.

data $Value\_10_x$ is set to zero. Then the system sends corresponding rainfall information to adjacent vehicle detectors range from the rainfall detector in 1 km.

The social media data is retrieved from the Police Broadcasting Service in Taiwan every ten minutes. The used parameters and the flowchart of social media data processing are listed in Table 3 and captured in Fig. 6. At the beginning, the system tries to access social data and will retry to access data in next time period if it fails to obtain the social media data of current time period. The parameter $Event\_type_x$ is utilized to check the type of traffic event x, such as traffic barrier, car accident, block, road construction. If the traffic event belongs to traffic congestion and traffic barrier, the reported event is mapped to the vehicle detectors. Therefore the reported event can be looked up to those vehicle detectors in the range of this event. Based on the comparison result, the system examines lane status and analyzes several reported information, e.g., average speed, the number of report events from different sections, and the location of report events.

TABLE 3. Parameter of social media data.

Parameter	Description
EV_URL	Social media file's URL
X	Total number of social media event
$Event\_type_x$	Type of report traffic event x
$Event\_location_x$	Location of reported traffic event x
$Event\_des_x$	Description of reported traffic event x
$Event\_date_x$	Date of reported traffic event x
$Event\_time_x$	Time of reported traffic event x
$Event\_source_x$	Source of reported traffic event x

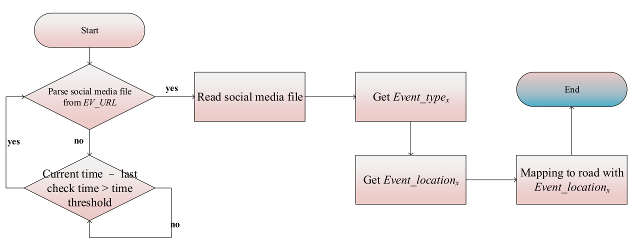


FIGURE 6. Flowchart of social media data processing.

D. CALCULATION OF CONGESTION LEVEL BY FUZZY

Vehicle data and weather data are sent to fuzzy module for road status analysis. The three stages of fuzzy theory are fuzzification, fuzzy inference and defuzzification. The used parameters in the proposed fuzzy module are listed in Table 4. In fuzzification stage, the membership function includes car speed, road density, rainfall and car volume. The fuzzy subsets of car speed with respect to low, normal and high are formulated as

TABLE 4. Parameter of fuzzy theory.

Parameter	Description
P_s	Car speed as the input of fuzzy logic control
P_d	Car density as the input of fuzzy logic control
P_r	Rainfall volume as the input of fuzzy logic control
P_v	Difference between historical and real-time car volume as the input of fuzzy logic control
a	The lower limit of membership fuction P_s, P_d
b	The middle limit of membership fuction P_s, P_d
c	The higher limit of membership fuction P_s, P_d
d	The lower limit of membership fuction P_r, P_v
e	The higher limit of membership fuction P_r, P_v

$$f_{low}(P_s) = \begin{cases} 1, & if\ 0 \le P_s < a \\ \dfrac{P_s - b}{b - a} + 1, & if\ a \le P_s < b \\ 0, & if\ P_s \ge b, \end{cases} \tag{3}$$

$$f_{normal}(P_s) = \begin{cases} \dfrac{P_s - b}{b - a} + 1, & if\ a \le P_s < b \\ \dfrac{a - P_s}{b - c} + 1, & if\ b \le P_s < c \\ 0, & otherwise, \end{cases} \tag{4}$$

$$f_{high}(P_s) = \begin{cases} 0, & if\ P_s < b \\ \dfrac{P_s - c}{c - b} + 1, & if\ b \le P_s < c \\ 1, & if\ b \le P_s. \end{cases} \tag{5}$$

The fuzzy subsets of road density with respect to low, normal and high are formulated as

$$f_{low}(P_d) = \begin{cases} 1, & if\ 0 \le P_d < a \\ \dfrac{P_d - b}{b - a} + 1, & if\ a \le P_d < b \\ 0, & if\ P_d \ge b, \end{cases} \tag{6}$$

$$f_{normal}(P_d) = \begin{cases} \dfrac{a - P_d}{b - c} + 1, & if\ a \le P_d < b \\ \dfrac{P_d - b}{b - a} + 1, & if\ b \le P_d < c \\ 0, & otherwise, \end{cases} \tag{7}$$

$$f_{high}(P_d) = \begin{cases} 0, & ifP_d < b \\ \dfrac{P_d - c}{c - b} + 1, & if\ b \le P_d < c \\ 1, & ifb \le P_d. \end{cases} \tag{8}$$

The fuzzy subsets of rainfall volume with respect to low and high are formulated as

$$f_{low}(P_r) = \begin{cases} 1, & if\ 0 \le P_r < d \\ \dfrac{P_r - e}{e - d} + 1, & if\ d \le P_r < e \\ 0, & if\ P_r \ge e, \end{cases} \tag{9}$$

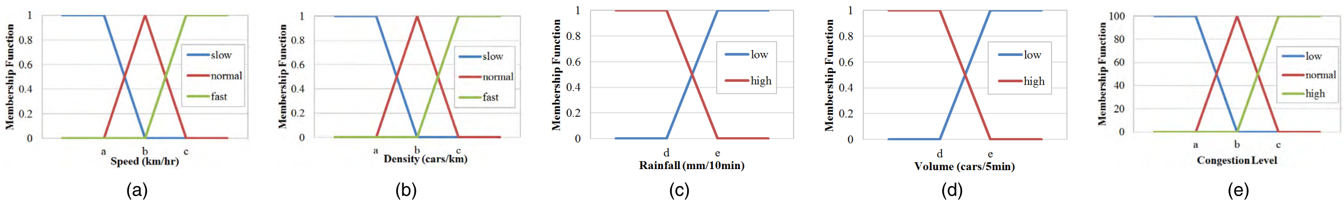


FIGURE 7. The membership function of proposed model. (a) Speed. (b) Density. (c) Rainfall. (d) Volume. (e) Congestion level.

$$f_{high}(P_r) = \begin{cases} 0, & if\ P_r < d \\ \dfrac{P_r - e}{e - d} + 1, & if\ d \leq P_r < e \\ 1, & if\ e \leq P_r. \end{cases} \tag{10}$$

The fuzzy subsets of car volume between historical and real-time with respect to low and high are formulated as

$$f_{low}(P_v) = \begin{cases} 1, & if\ 0 \leq P_v < d \\ \dfrac{P_v - e}{e - d} + 1, & if\ d \leq P_v < e \\ 0, & if\ P_v \geq e, \end{cases} \tag{11}$$

$$f_{high}(P_v) = \begin{cases} 0, & if\ P_v < d \\ \dfrac{P_v - e}{e - d} + 1, & if\ d \leq P_v < e \\ 1, & if\ e \leq P_v. \end{cases} \tag{12}$$

The obtained membership functions of car speed, road density, rainfall and car volume are shown as Fig. 7. In fuzzy inference stage, the fuzzy inference is conducted based on the fuzzy rules in fuzzy database. The fuzzy rule is defined as

$$R^{(l)}:\ if\ x\ is\ A_i^l\ then\ y\ is\ B^l, \tag{13}$$

where $R^{(l)}$ is the l-th rule, x and y are the input and output of fuzzy module. The parameters A_i^l and B^l are fuzzy sets. The proposed fuzzy inference is based on minimum inference mechanism, which is defined as

$$B'(y) = \max_{1 \leq l \leq m} \left[A_1^l(x_1) \cdot A_2^l(x_2) \cdot B^l(y) \right]. \tag{14}$$

In the last stage, the proposed defuzzification is based on the center of area method and defined as

$$y = \frac{\int_Y yB(y)dy}{\int_Y B(y)dy}. \tag{15}$$

According the three stages, the congestion level of a road section can be quantified to 0 to 100.

E. PREDICTION OF ROAD SPEED BY SVM

The data format should be defined thus the car speed data can be applied to SRHTCP's prediction process. The data format of SRHTCP is listed as Table 5. The label stands for the category of car speed per hour. First, the label is used to classify car speed. The index represents the dimension of SRHTCP's training set as well as the data's feature value. The value stands for the realistic value of the dimension.

In the work, the car speeds in previous three time periods are used to be the dimension of training data in SRHTCP. The parameter S_t represents the car speed in time t, and S_{t-1}, S_{t-2} and S_{t-3} are the car speed of time $t-1$, $t-2$ and $t-3$ respectively. The S_{t-1}, S_{t-2} and S_{t-3} are used to conduct the car speed per hour of S_t. Therefore the travel time of S_t can be calculated by three previous time periods S_{t-1}, S_{t-2} and S_{t-3}. The flowchart of data processing is captured in Fig. 8.

TABLE 5. Data format of SRHTCP.

[Label]	[Index1]:[Value1]	[Index2]:[Value2]	[Index3]:[Value3]
[Label]	[Index1]:[Value1]	[Index2]:[Value2]	[Index3]:[Value3]
[Label]	[Index1]:[Value1]	[Index2]:[Value2]	[Index3]:[Value3]
...	...	...	...

First of all, parsing input variable initializes the value of $feature\_amount_g$. It means that how many features are needed to describe each label in the calculation of SRHTCP. In addition, the number of support vector h is initialized to zero. Then, features are retrieved from each data file so that SRHTCP examines where the number of feature is larger than the $feature\_amount_g$ author defined or not. If the number of features is less than the defined $feature\_amount_g$, it means the number of features is insufficient so that the label's feature number should be increased continuously. When feature number exceeds the $feature\_amount_g$, SRHTCP starts to retrieve next feature until there is no more support vectors.

After retrieving features, the proposed system classifies training data into several groups based on the parameters user defined. The system utilizes one arbitrary parameter to train and predict the accuracy of this training data. Then it uses another parameter to execute data training and others are used to predict accuracy. After the training process, one of the parameters is selected to be the optimal solution and utilized in support vectors for data training. After training process, the trained model predicts data sets and obtains final result.

IV. EXPERIMENTAL RESULT AND ANALYSIS

In this section, congestion prediction with big data for freeway traffic is implemented. Based on the concept of big data, Apache Storm is used to implement platform that collects

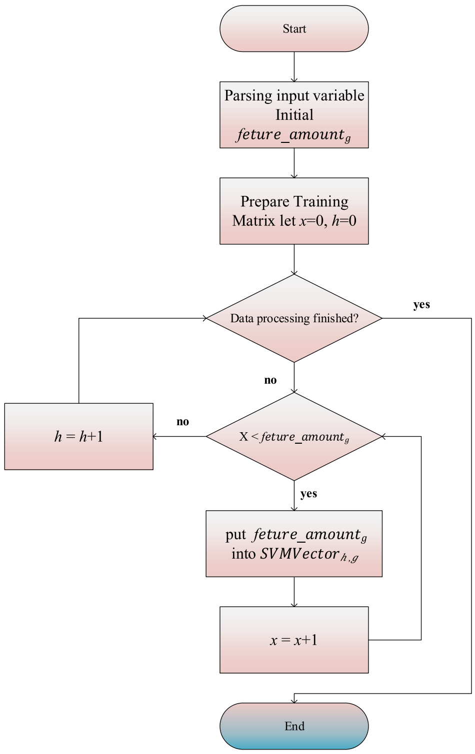


FIGURE 8. The flowchart of data processing.

TABLE 6. Specification of XenServer.

Hardware/Software	Specification
CPU	AMD Opteron(tm) Processor 6172
Memory	16382 Mb
Network interface card	Intel® 82576 Gigabit Ethernet Controller
Network interface	1000 Base TX
Operating system	Xen server 6.2

traffic, weather and social data. The congestion level of freeway is analyzed by the proposed fuzzy model and the car speed of next time period is predicted by the proposed SRHTCP model.

A. EXPERIMENT SET UP AND PARAMETERS

The experiments in this paper are implemented by the VMs in XenServer. The hardware and software of used environment in XenServer are captured in Table 7. In the constructed Apache Storm platform, there is a master node, a ZooKeeper node and a slave node. Some of used hardware and software resources in the master node, ZooKeeper node and slave node are the same with used specification in XenServer, i.e., CPU, network interface card and network interface. In master, ZooKeeper and slave node, the used memory is 4096 Mb. In master node, the operating system (OS) version is Ubuntu 14.04.2 LTS and the OS kernel is 3.13.0-55-generic. The operating system of ZooKeeper node is XenServer 6.2. Both master node and slave node are equipped with Apache Storm in version 0.9.0.1.

B. MEASUREMENT OF PREDICTION ACCURACY

In the paper, the prediction accuracy is measured by the Mean Absolute Relative Error (MARE) and Mean Square Error (MSE) methods. The MARE is calculated by

$$\mathrm{MARE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|a_i - b_i|}{a_i}, \tag{16}$$

where a_i is the prediction result and b_i is the observation data. The MSE is calculated by

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(a_i - b_i)^2, \tag{17}$$

where a_i and b_i is the value of prediction and observation data respectively.

The MARE and MSE are familiar methods to evaluate the difference between prediction and reality [33]. The MARE is a percentage and the MSE is a value. The MARE percentage represents the difference between prediction result and observation data. A lower MARE percentage stands for the higher prediction accuracy, and vice versa. However, the MARE is hard to explore the difference between prediction and observation when the error difference is small. The MSE is more useful to enlarge error difference so that the prediction error will be more obvious. In other words, the prediction accuracy and inaccuracy are easier to be explored. A lower MSE implies that the higher prediction accuracy, and vice versa. Both methods are used to evaluate prediction accuracy in the paper.

C. EXPERIMENT 1: REAL-TIME TRAFFIC ANALYSIS BASED ON TRAFFIC REPORTS FROM DRIVERS

In this experiment, the real-time traffic reports from Police Broadcasting Service in Taiwan are collected and matched with the information retrieved by those vehicle detectors where in the same region. As a result, the matched information is more valuable than the information from vehicle detectors without matching with traffic reports. The information reported by drivers or passengers can reflect the congestion level at that time. We investigate the number of traffic reports under different car speed, which is captured in Fig. 9. It can be observed that drivers report traffic jam more times when the car speed equals to 50 to 60 and 90 to 100. It implies that drivers not only report traffic jam when congestion happened, sometimes drivers but also report traffic jam when there is a slight congestion.

In addition, we also investigate the car speed of different counties and cities in Taiwan, which is captured in Fig. 10. The first four counties and cities are in north Taiwan, the last three counties are in south Taiwan, and others are in the middle of Taiwan. It can be observed that there are more traffic report times in north Taiwan and fewer reports in middle and south Taiwan. This result is mainly attributed to the fact that drivers in north Taiwan are used to listen the Police Broadcasting Service radio. Thus there are more traffic reports in north than other counties and cities in Taiwan.

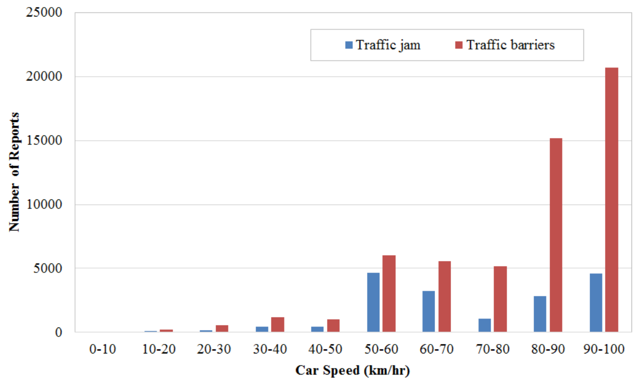


FIGURE 9. Number of reports under different car speed.

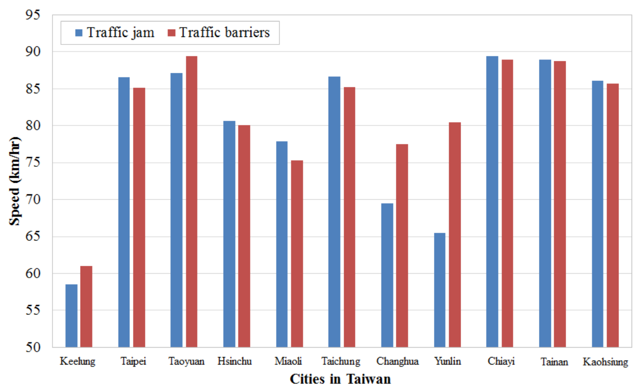


FIGURE 10. Number of reports in different counties and cities.

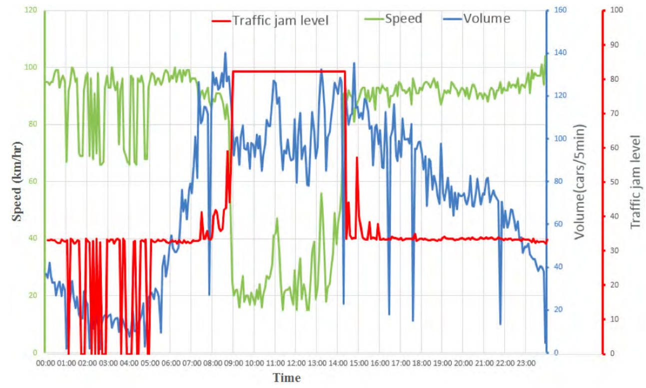


FIGURE 11. Traffic jam level in a non-rainy day.

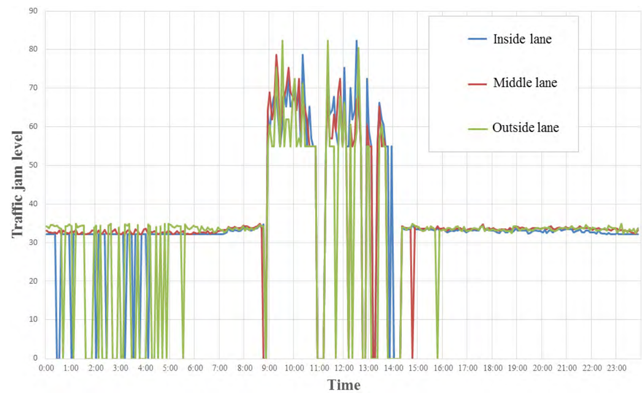


FIGURE 12. Traffic jam level of different lanes.

D. EXPERIMENT 2: REAL-TIME TRAFFIC ANALYSIS OF A NON-RAINY DAY

In this experiment, real-time traffic in non-rainy day is analyzed to examine whether weather influences traffic or not. The traffic jam level of a non-rainy day is captured in Fig. 11. The observation site is located in Hsinchu interchange, and the date is May 1, 2015. It can be observed that from 9 am to 2 pm, the higher vehicle volume and lower car speed leads to the higher traffic jam level. This result is mainly attributed to the fact that Hsinchu is an industrial area and from 9 am to 11 am is the commute time. In addition, from 11 am to 2 pm is the lunchtime, Hsinchu industrial area is lack of good restaurant so that engineers are used to drive their cars for lunch.

In Fig. 12, the traffic jam level of different lanes is captured. It can be observed that the traffic jam level from 9 am to 2 pm is higher than other time periods. The commute time and lunchtime result in the higher congestion level. On the other hand, it can be observed that the congestion level of inner lane and middle lane is steadier than that of outside lane. This is attributed to the fact that cars in outside lane have higher probability of leaving freeway and blocked by traffic lights.

E. EXPERIMENT 3: REAL-TIME TRAFFIC ANALYSIS OF A RAINY DAY

In this experiment, real-time traffic of a rainy day is analyzed. In Fig. 13, the traffic data is captured by the vehicle detectors from Taipei to Sanchong interchange, and the observation data is June 14, 2015. It can be observed that the rain reaches about 3 mm every ten minutes at 3:30 pm and slightly decreases to 1.5 mm every ten minutes at 4 pm and 5 pm. In general, rainy days affect driver's sight and vision so that drivers drive slowly. The car speed is obviously lower at 3 pm than other time periods due to the heavy rain. In addition, it can be observed that the vehicle volume at 3 pm is lower than other time periods around 80 cars every five minutes but the car speed is still low. This result is mainly attributed to the fact that heavy rain leads to slower driving.

In Fig. 14, we also investigate the traffic jam level of the same rainy day. The traffic jam level in Fig. 13 is evaluated and obtained by the proposed fuzzy model. It can be observed that the traffic jam level at 3 pm is higher than time, the same phenomenon happened at 4:30 pm. Both results mainly attributed to the fact that heavy rain results in drivers drive slowly, thus the slower car speed leads to higher traffic jam level. The fuzzy inference of proposed model is validated by the results.

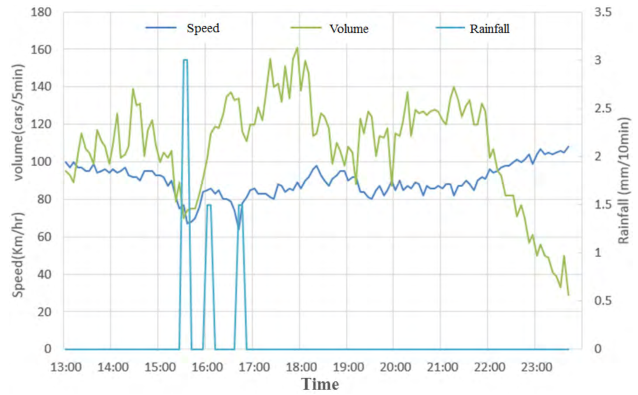


FIGURE 13. Traffic analysis of a rainy day.

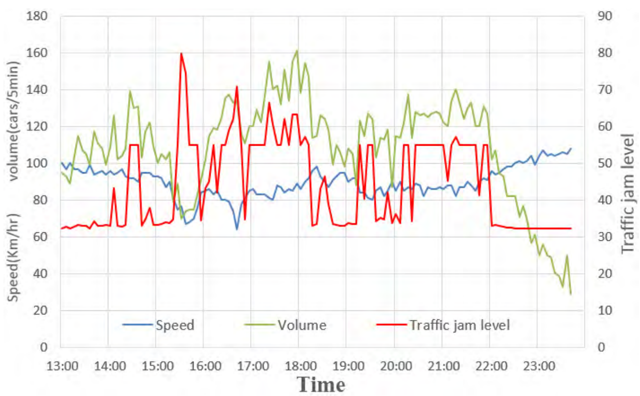


FIGURE 14. Traffic jam level of a rainy day.

F. EXPERIMENT 4: CAR SPEED PREDICTION BY SRHTCP AND EWMA

In this experiment, real-time car speed is predicted by the proposed SRHTCP method and the Exponentially Weighted Moving Average (EWMA) method. The real-time car speed of current, five minute and ten minute are used to be the training data. We propose SRHTCP model to explore the feature values of different car speed so that the car speed in next time period can be predicted accurately. The training data collected from May 1, 2015 to June 9, 2015. The prediction result records from June 10 to 16, 2015. The prediction results of SRHTCP and EWMA are captured in Table 8. Note that the lower MARE and MSE stand for the higher prediction accuracy. It can be observed that the proposed SRHTCP forecasts the car speed in next week accurately, which the MARE and MSE is less than 4.27% and 88.89 respectively.

On the other hand, the SRHTCP method is compared with the Exponentially Weighted Moving Average (EWMA) method. The EWMA method is also used to forecast the car speed of next time period, which is calculated by

$$\mathrm{EWMA}_n = a * \mathrm{price} + (1 - a) * \mathrm{EWMA}_{n-1}, \qquad (18)$$

where EWMA_n is the prediction result of next time period, a is a weighted value and price is current value, and EWMA_{n-1} is the prediction result of previous time period. Note that we set the weight a equals to 0.125, which is a common value in computer networks. The used training data in EWMA method is the same with the training data in SRHTCP method. It can be observed that the prediction accuracy of EWMA method is worse than that of the proposed SRHTCP, no matter in what kind of estimation criteria, i.e., MARE and MSE. Thereby we can say that the proposed SRHTCP method is superior to the EWMA method in terms of prediction accuracy.

TABLE 7. Prediction result of SRHTCP and EWMA.

Date	SRHTCP MARE	EWMA MARE	SRHTCP MSE	EWMA MSE
June 10	2.96 %	4.39 %	29.99	39.01
June 11	4.14 %	4.24 %	38.10	37.80
June 12	4.15 %	4.22 %	41.03	37.25
June 13	2.32 %	3.32 %	19.07	27.30
June 14	4.27 %	4.18 %	88.89	93.25
June 15	3.87 %	6.96 %	55.59	106.44
June 16	1.46 %	3.83 %	12.06	31.56
Average	3.31 %	4.45 %	40.68	53.20

TABLE 8. Prediction result of case 1 and case 2.

Date	Case 1 MARE	Case 2 MARE
June 10	3.85%	3.02%
June 11	5.14%	5.18%
June 12	4.15%	4.66%
June 13	3.44%	2.95%
June 14	4.50%	5.52%
June 15	7.31%	6.00%
June 16	3.78%	0.39%
Average	4.60%	3.93%

G. EXPERIMENT 5: CAR SPEED PREDICTION BY USING THE TRAINING DATA IN DIFFERENT TIME PERIODS

In this experiment, we use the proposed SRHTCP model to predict the car speed of 30 minutes later. Two sets of training data are used and named as case 1 and case 2. In case 1, the SRHTCP model is trained by the data in current, previous 5 and 10 minutes. In case 2, training data is obtained from the data in current, previous 10, 20 and 30 minutes. The SRHTCP retrieves feature values from these two training data sets and forecasts the car speed of 30 minutes later. The training data of this experiment starts from May 1 to June 9, 2015. The prediction results of SRHTCP in case 1 and case 2 are captured in Table 9. It can be observed that the SRHTCP model yields the higher prediction accuracy in case 2 than case 1. The average MARE value of SRHTCP in case 1 and case 2 is 4.6% and 3.93% respectively. The proposed SRHTCP model improves 14.57% prediction accuracy in case 2 compared with case 1. This result is mainly attributed to the fact that the SRHTCP can find feature value more accurately in case 2 because there are more training data.

V. CONCLUSIONS

Unlike existing literatures used batch method to predict car traffic, we utilize Apache Storm platform to achieve real-time traffic prediction. The constructed platform integrates different kinds of open data, i.e., traffic data from Taiwan Area National Freeway Bureau, weather data from Central Weather Bureau, and social media data from Police Broadcasting Service. By analyzing the great quantity of traffic data, we found that there are two sorts of traffic pattern in Taiwan, i.e., weekdays from Monday to Thursday and weekend from Friday to Sunday. We analyzed social media data and found that drivers in inner line inform traffic jam report when car speed is lower than 60 km per hour. In addition, drivers in south Taiwan inform traffic jam when car speed is lower than 90 km per hour. It implies that drivers in south Taiwan have less tolerance of car speed. In experiments, we not only utilized fuzzy theory to analyze real-time traffic and congestion level but also proposed SRHTCP model to forecast the car speed of next time period. It has been shown that the SRHTCP model is superior to the EWMA method in terms of prediction accuracy no matter in MARE or MSE analysis. In the future, we will try to verify the used open data sets with t-test method.

REFERENCES

[1] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big data analytics: A survey," *J. Big Data*, vol. 2, no. 21, pp. 1–32, Oct. 2015.

[2] D. Laney, "3D data management: Controlling data volume, velocity and variety," META Group, Stamford, CT, USA, Tech. Rep. 949, Feb. 2001.

[3] P. Chandarana and M. Vijayalakshmi, "Big data analytics frameworks," in *Proc. CSCITA*, Mumbai, India, 2014, pp. 430–434.

[4] OICA. (May 1, 2018). *Organisation Internationale des Constructeurs d'Automobiles*. [Online]. Available: http://www.oica.net/

[5] (May 1, 2018). *Ministry of Transportation and Communications R.O.C.* [Online]. Available: http://www.motc.gov.tw/ch/index.jsp

[6] *Apache Storm*. (Dec. 24, 2015). [Online]. Available: http://storm.apache.org/

[7] (Jul. 19, 2017). *Apache Storm Rationale Version 1.1.0*. [Online]. Available: http://storm.apache.org/releases/current/Rationale.html

[8] (Jul. 19, 2017). *DATA.GOV.TW*. [Online]. Available: http://data.gov.tw/

[9] (Jul. 19, 2017). *Tisvcloud.Freeway.Gov.Tw*. [Online]. Available: http://tisvcloud.freeway.gov.tw/

[10] (Jul. 19, 2017). *OPEN DATA*. [Online]. Available: http://opendata.cwb.gov.tw/index

[11] (Jul. 19, 2017). *Police Broadcasting Service of Taiwan*. [Online]. Available: http://rtr.pbs.gov.tw/pbsmgt/RoadAll.html

[12] B. D. Greenshields, "A study of traffic capacity," in *Proc. Highway Res. Board*, vol. 14, 1935, pp. 448–477.

[13] A. Nakayama *et al.*, "Detailed data of traffic jam experiment," in *Traffic & Granular Flow*. Berlin, Germany: Springer, 2009, pp. 389–394.

[14] University of Bristol. (Jul. 19, 2017). *Phantom Traffic Jams*. [Online]. Available: http://www.bristol.ac.uk/news/2010/6948.html

[15] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338–353, Jun. 1965.

[16] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. ACM 5th Annu. Workshop Comput. Learn. Theory (COLT)*, Pittsburgh, PA, USA, Jul. 1992, pp. 144–152.

[17] F.-H. Tseng, X. Chen, L.-D. Chou, H.-C. Chao, and S. Chen, "Support vector machine approach for virtual machine migration in cloud data center," *Multimedia Tools Appl.*, vol. 74, no. 10, pp. 3419–3440, May 2015.

[18] C.-Y. Chen, K.-D. Chang, and H.-C. Chao, "Transaction-pattern-based anomaly detection algorithm for IP multimedia subsystem," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 1, pp. 152–161, Mar. 2011.

[19] W. Wu and H. Zhou, "Data-driven diagnosis of cervical cancer with support vector machine-based approaches," *IEEE Access*, vol. 5, pp. 25189–25195, Oct. 2017.

[20] S. U. Jan, Y. D. Lee, J. Shin, and I. Koo, "Sensor fault classification based on support vector machine and statistical time-domain features," *IEEE Access*, vol. 5, pp. 8682–8690, May 2017.

[21] M. Milojevic and V. Rakocevic, "Distributed vehicular traffic congestion detection algorithm for urban environments," in *Proc. IEEE Veh. Netw. Conf. (VNC)*, Boston, MA, USA, Dec. 2013, pp. 1–6.

[22] Y. Cheng, M. Zhang, and D. Yang, "Automatic incident detection for urban expressways based on segment traffic flow density," *J. Intell. Transp. Syst.*, vol. 19, no. 2, pp. 205–213, Jan. 2015.

[23] H. Ji, A. Xu, X. Sui, and L. Li, "The applied research of Kalman in the dynamic travel time prediction," in *Proc. 18th Int. Conf. Geoinformat.*, Beijing, China, Jun. 2010, pp. 1–5.

[24] H. Feng, C. Liu, Y. Shu, and O. W. W. Yang, "Location prediction of vehicles in VANETs using a Kalman filter," *Wireless Pers. Commun.*, vol. 80, no. 2, pp. 543–559, Jan. 2015.

[25] H.-F. Wang, "A freeway automatic incident detection algorithm using grey prediction," M.S. thesis, Dept. Comput. Sci. Inform. Eng., Nat. Central Univ., Taoyuan, Taiwan, 2003.

[26] K.-W. Kuo, "A hybrid travel-time prediction approach based on macroscopic and microscopic methodologies," M.S. thesis, Dept. Comput. Sci. Inform. Eng., Nat. Central Univ., Taoyuan, Taiwan, 2013.

[27] D. C. Li, L.-D. Chou, L.-M. Tseng, Y.-M. Chen, and K.-W. Kuo, "A bipolar traffic density awareness routing protocol for vehicular ad hoc networks," *Mobile Inf. Syst.*, vol. 2015, Sep. 2015, Art. no. 401518, doi: 10.1155/2015/401518.

[28] W. Yang, X. Liu, L. Zhang, and L. T. Yang, "Big data real-time processing based on storm," in *Proc. 12th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, Melbourne, VIC, Australia, Jul. 2013, pp. 1784–1787.

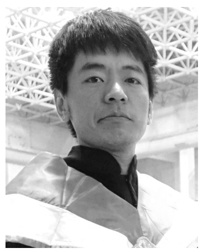
[29] T. Chardonnens, P. Cudre-Mauroux, M. Grund, and B. Perroud, "Big data analytics on high velocity streams: A case study," in *Proc. IEEE Int. Conf. Big Data*, Silicon Valley, CA, USA, Oct. 2013, pp. 784–787.

[30] *Twitter*. (Dec. 24, 2015). [Online]. Available: https://twitter.com/

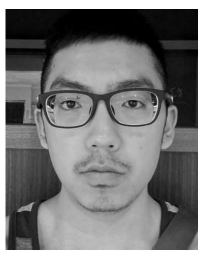
[31] *Bitly*. (Dec. 24, 2015). [Online]. Available: https://bitly.com/

[32] A. Bifet and G. De F. Morales, "Big data stream learning with SAMOA," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Shenzhen, China, Dec. 2014, pp. 1199–1202.

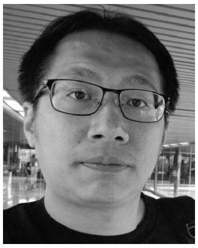
[33] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature," *Geosci. Model Develop.*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014.



FAN-HSUN TSENG (S'12–M'18) received the Ph.D. degree in computer science and information engineering from National Central University, Taiwan, in 2016. He is currently an Assistant Professor with the Department of Technology Application and Human Resource Development, National Taiwan Normal University, Taipei, Taiwan. His research interests include mobile edge computing, 5G mobile networks, and artificial intelligence. He received the 2018 Young Scholar Fellowship Program from the Ministry of Science and Technology, Taiwan, for his dedication to research of engineering and technologies. He has served as an Associate Editor-in-Chief of the *Journal of Computers* and an Associate Editor of *Human-Centric Computing and Information Sciences*.



JEN-HAO HSUEH received the M.S. degree in computer science and information engineering from National Central University, Taoyuan, Taiwan, in 2015. His research interests include big data, vehicle ad-hoc network, fuzzy theory, and support vector machine.



CHIA-WEI TSENG (S'17) received the M.S. degree in computer science and information engineering from National Dong Hwa University, Hualien, Taiwan, in 2006. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Information Engineering, National Central University, Taiwan. His research interests include software-defined networking, vehicle ad-hoc network, and IPv6 protocol.

YAO-TSUNG YANG (S'10) received the M.S. degree in computer science and information engineering from National Central University, Taoyuan, Taiwan, in 2008, where he is currently pursuing the Ph.D. degree with the Department of Computer Science and Information Engineering. His research interests include software-defined networking, vehicle ad-hoc network, and network management.

HAN-CHIEH CHAO (SM'04) received the M.S. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1989 and 1993, respectively. He is currently a Professor with the Department of Electrical Engineering, National Dong Hwa University, where he also serves as the President. He is also with the Department of Computer Science and Information Engineering and the Department of Electronic Engineering, National Ilan University, Taiwan, the College of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan, China, and the Fujian University of Technology, Fuzhou, China. He is a fellow of IET (IEE) and a Chartered Fellow of the British Computer Society. He serves as the Editor-in-Chief for the Institution of Engineering and Technology Networks, the *Journal of Internet Technology*, the *International Journal of Internet Protocol Technology*, and the *International Journal of Ad Hoc and Ubiquitous Computing*.

LI-DER CHOU (M'95) received the M.S. and Ph.D. degrees in electronic engineering from the National Taiwan University of Science and Technology, Taiwan, in 1991 and 1995, respectively. He is currently a Distinguished Professor with the Department of Computer Science and Information Engineering and the Director of Computer Center, National Central University, Taiwan. He was a Director of the Board of Taiwan Network Information Center. He was also the Deputy Director General of the National Center for High-performance Computing, Taiwan, from 2013 to 2016. He is the holder of five U.S. and 16 Taiwan invention patents. His research interests include SDN/NFV/SFC, vehicular networks, network management, broadband wireless networks, and Internet services. He has published over 200 papers in these areas. He was a recipient of seven best paper awards and four excellent paper awards from international and domestic conferences. He was also a recipient of two Gold Medal Awards and four Silver Medal Awards in international invention shows held in Geneva, Moscow, London, and Taipei.