World Academy of Science, Engineering and Technology
International Journal of Architectural and Environmental Engineering
Vol:11, No:10, 2017

Integration of Big Data to Predict Transportation for Smart Cities

Sun-Young Jang, Sung-Ah Kim, Dongyoun Shin

***Abstract*—**The Intelligent transportation system is essential to build smarter cities. Machine learning based transportation prediction could be highly promising approach by delivering invisible aspect visible. In this context, this research aims to make a prototype model that predicts transportation network by using big data and machine learning technology. In detail, among urban transportation systems this research chooses bus system. The research problem that existing headway model cannot response dynamic transportation conditions. Thus, bus delay problem is often occurred. To overcome this problem, a prediction model is presented to fine patterns of bus delay by using a machine learning implementing the following data sets; traffics, weathers, and bus statues. This research presents a flexible headway model to predict bus delay and analyze the result. The prototyping model is composed by real-time data of buses. The data are gathered through public data portals and real time Application Program Interface (API) by the government. These data are fundamental resources to organize interval pattern models of bus operations as traffic environment factors (road speeds, station conditions, weathers, and bus information of operating in real-time). The prototyping model is designed by the machine learning tool (RapidMiner Studio) and conducted tests for bus delays prediction. This research presents experiments to increase prediction accuracy for bus headway by analyzing the urban big data. The big data analysis is important to predict the future and to find correlations by processing huge amount of data. Therefore, based on the analysis method, this research represents an effective use of the machine learning and urban big data to understand urban dynamics.

***Keywords*—**Big data, bus headway prediction, machine learning, public transportation.

I. Introduction

IN order to maintain and to enhance urban systems, intelligence urban operation system are crucial. The urban facilities are generally classified as architectural facilities, transportation facilities, and flat facilities [1]. Above all, transportation facilities are one of the most dynamic elements in the cities fundamentally integrated in urban activities (such as producing, living, entertaining, etc.) functionally.

Among public transportations, bus is a crucial urban transportation system to citizens in daily lives. The recent public transportation systems have been improved the convenience of use by applying various ICTs (Information and Communications Technology). Citizens could be received information about bus locations, waiting times, and remaining seats from the bus information system (BIS) in real-time. By using the BIS, the practical use of the bus information is smarter than in the past. Before using the BIS, we relied on getting information from the attached timetable and route information in the bus station. To construct such BIS and enhance the convenience of using public transportations through the system are also among the efforts to build the smart city [2].

Citizens are enabling to gain more information in the use of bus. When problems of bus services occurred, however, the problems are becoming burdens to the citizens. Citizens are subordinated to the problems regardless of the convenience of using information. For example, interval problems occur because of weather conditions or fails of interval control with buses. These problems arise because the current bus systems do not respond to the circumstances of city immediately. The bus companies endeavor to deal with the problems coming up frequently based on the driving experiences in a long time. However, the ways to treat the problems lack of elaborate analysis, and tend to be determined by company or government by political reasons.

To resolve the problems, the local governments and transportation research institutions investigate and implement a flexible headway to the demands and traffic conditions [3], [4]. These researches build a model to carry out a flexible headway based on the various statistics about urban lifestyles, transportation statuses, and operation records of the past. This way also provides models in context using a sort of accumulated data. Recent technological advances, for example the use of big data and machine learning, improve the system efficiency. These technologies are already applied to improve transportation systems in many countries, and required to active use to construct the intelligent transportation systems [5]. Based on the abilities of the data analysis and utilization, the transportation system of smart city should be able to have flexibility to respond to urban conditions (ex. traffic congestion, large scale event, weather, demand, etc.).

II. Research Objectives

The objectives of research are to present implications about using urban big data and machine learning. It is a method to construct the transportation network of smart city that have a flexible system to respond the urban context changes. This research addresses a problem that existing headway model is usually a fixed type and thus difficult to respond urban context in real-time. Therefore, using a machine learning tool and urban big data in traffics, weathers, and bus statues, this research provides a flexible headway model based on the urban big data through prototyping model by predicting bus delay and

Sun-Young Jang, Ph.D. student, and Professor Sung-Ah Kim are with the Department of Architecture, Sungkyunkwan University, Suwon, Republic of Korea (e-mail: abyme1204@ skku.edu, sakim@ skku.edu).

Dr. Dongyoun Shin is with the Department Information Architecture, ETH Zurich, Switzerland (e-mail: dongyoun79@gmail.com).

analyzing the result. The presented prototype model is built with real-time operating data of an actual bus. The data is collected through open data of public data portals [15]-[17], analyzed data, and real time API by the government. The data is basic resources to compose interval pattern models of bus operations by elements such as weathers, road speeds, station conditions, and bus information of operating in real-time. The model is built by the machine learning tool (RapidMiner Studio) and tested to predict bus delays in various circumstances. Finally, on this research shows the experiment and give a discussion how transportation systems responding to urban conditions and possibilities to promote the citizens' convenience for the urban efficiency.

III. Literature Reviews on the Headway and Transportation Research Works Using Big Data

The headway of city bus is a crucial element about the convenience and confidence of the bus operation in evaluation of the bus service level. And it is involved in the cost and operation efficiency of the bus company. Likewise, the local governments, it is also important to set the headway of city bus with considering the various stances and environmental factors because of their significant financial aid. Therefore, the previous research studies on the bus headway have purposes to improve the bus service level on the side of the citizens [6]-[8]. These researches aim to calculate the number of bus operation or to decide optimal bus routes. And It has tried to find an effective head ways on the side of the whole city or the company [9]-[11]. The researches, to present operation methods of flexible headway by analyzing based on the previous cumulated traffic circumstances [3] or strategies for the effective headway by reflecting bus context in real-time [12], almost apply the methodologies of mathematical statistics based on the existing data. Recently, the cases of applying the big data are increased. Also, the mentioned researches are also in company with the big data as statistics. However, the current studies in the big data domain include more active and divers use of data spectrum. It predicts complex relations that people could not even consider in generating the massive amounts of information in seconds. A case of improvement in the traffic system applying the big data of the USA collects and analyzes the real-time data with public-private partnership.

Especially, the private sectors are growing a role because data qualities are progressed and collected the massive traffic data in real-time through the diverse collective ways (extraction from GPS data, Bluetooth, and camera of smartphone, and signal of mobile device). The Seoul City deduces the optimal bus routes and the headways for the public transportation service by analyzing patterns in the big data about the traffic demands and floating populations.

This research is not intended to solve the problems by considering a model of the noble bus headway with considering every situation. The mathematical algorithms to set and adjust the bus headway are significantly complicated and involved many factors influencing mutually. However, this research contributes to mitigate the problems with raising the accuracy to fulfill the bus headway by analyzing and reflecting the urban big data. The analysis of big data has a signification to predict the future by processing massive amount of data with complicated relations. Based on the analysis method, therefore, this research presents an effective use of the urban big data and machine learning to solve the problems by using the several types of urban data related to bus delay cases.

IV. A Prediction Model of Bus Delay Using Real-Time Bus Data

A. The Urban Data Gathering and Extraction

This research used four types of data; weather conditions, road speeds, bus locations, and bus stations. The data are collected at the three different sources, open data portal of the government [15], the Gyeonggi bus information system (GBIS) of the local government of the Gyeonggi province [16], and open weather data portal of the national weather center [17]. The data collecting period are September 18 to 27, 2017. And it is focused on the morning (between 7 and 9am) and evening (between 6 and 8 pm) rush hour.

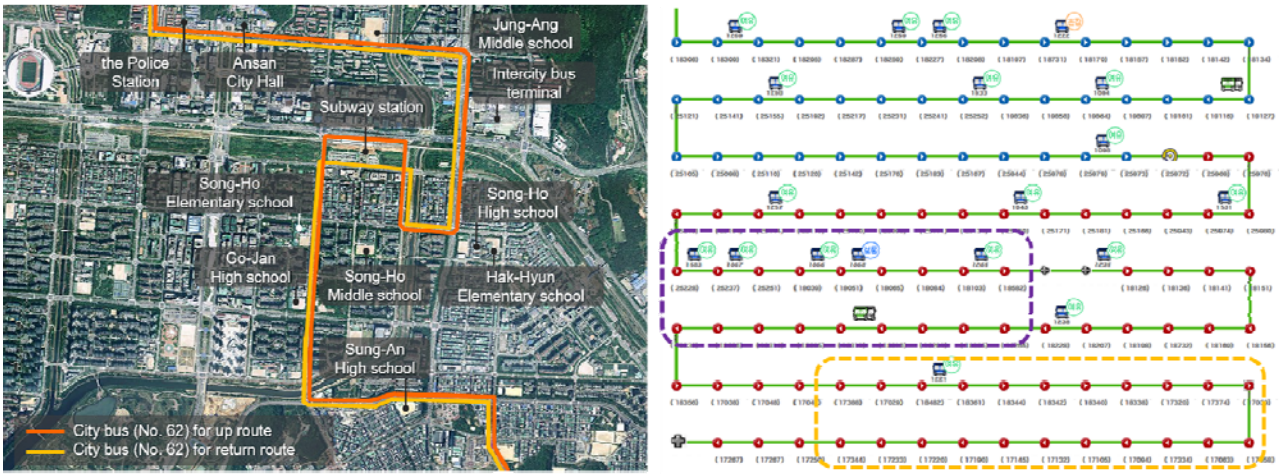

Fig. 1 The target route of data collection and the situation of density (map: naver portal site, real-time bus information: GBIS)

TABLE I
META DATA OF THE DATA-SET FOR PREDICTING BUS DELAY

Feature	Value type	Role	Description	Value
examples-time	date_time	attribute	example	numeric value
W-precipitation	real	attribute	Precipitation (mm)	numeric value
W-type	integer	attribute	precipitation type (code)	0(none), 1(rain), 2(sleet), 3(snow)
W-temp	real	attribute	temperature (℃)	numeric value
T-speed	integer	attribute	load speed between stations (km/h)	numeric value
B-stop	integer	attribute	station ID (seven-digit serial number)	numeric value
B-predictT1	integer	attribute	prediction time of the first arrival bus (minutes)	numeric value
B-predictT2	integer	attribute	prediction time of the second arrival bus (minutes)	numeric value
B-interval	integer	attribute	interval between first and second arrival buses (minutes)	numeric value
IntervalType2	binominal	label	under 10 minutes: 1/ over 11 minutes (delay): 0	0(Yes), 1(No)

The target on this research is the city bus number 62 which stops by main city area of Ansan city. The city bus (No. 62) has 73 bus stops for up route and 77 stops for return route. The data is collected for all stations and mainly focused on down 17 stations (5km) passing through major city center, such as schools (elementary: 4, middle: 4, high: 3, university: 1), large size churches, the City Hall, the police station, the intercity bus terminal, and subway stations. Therefore, the target stops currently has irregular bus intervals off the time schedule (Fig. 1). The weather data contains precipitation type (none/ rain/ sleet/ snow), precipitation (mm), and temperature (℃). The road speed data are collected each section between the stations from the GBIS. The bus data are collected in two-minute interval at each station from the open API of the GBIS and route information from the open data portal. The data include station ID, route ID, location data of first and second arrival buses, arrival prediction time, low floor bus, and bus number.

B. Composition of Bus Delay Prediction Model

The bus delay prediction model is developed by using machine learning prototyping tool (RapidMiner Studio). The model predicts bus delay using the following circumstance data sets, patterns of accumulated bus driving data, weather data, road speed, etc. The city bus number 62 specifies their bus headway as 5 to 10 minutes on weekdays. This experiment counts as "interval fail" when the bus interval is under 3 minutes, and over 11 minutes.

The prediction model presents the results as follows; 1) bus delay prediction, 2) influence factors through calculating weight values of the conditional elements influencing the bus delay, 3) analysis through the correlation matrix. The Fig. 2 represents the model composition and operating process.

Process 1– The training data set is loaded using Read Excel operator. Training data set and test data set include meta data of 10 types addressed in Table 1. The number of examples abstracted for training are different in every test. Multiply Data operator acts to use training data to weight calculating (Process 2), probability calculating (Process 3), and Correlation Matrix (Process 5) concurrently. Process 2– This phase deduces the weight value which influence on the label. The Calculate Weights operator calculates the relevance of attributes. Process 3– Training of delay prediction model utilizes the Naïve Bayes operator. The operator generates a Naïve Bayes classification model based on probability. This method judges the results by probabilistic classification of positive and negative. Process 4- The trained model predicts bus delay of test data set. The prediction values indicate 0 (Yes) or 1 (No) and deduce confidences about the judgement. The accuracy of prediction result is produced by comparing real delay data. Process 5- The Correlation Matrix operator is used to analyze the influence factors. This phase judges influence factors to label and presents correlations between the independent variables.

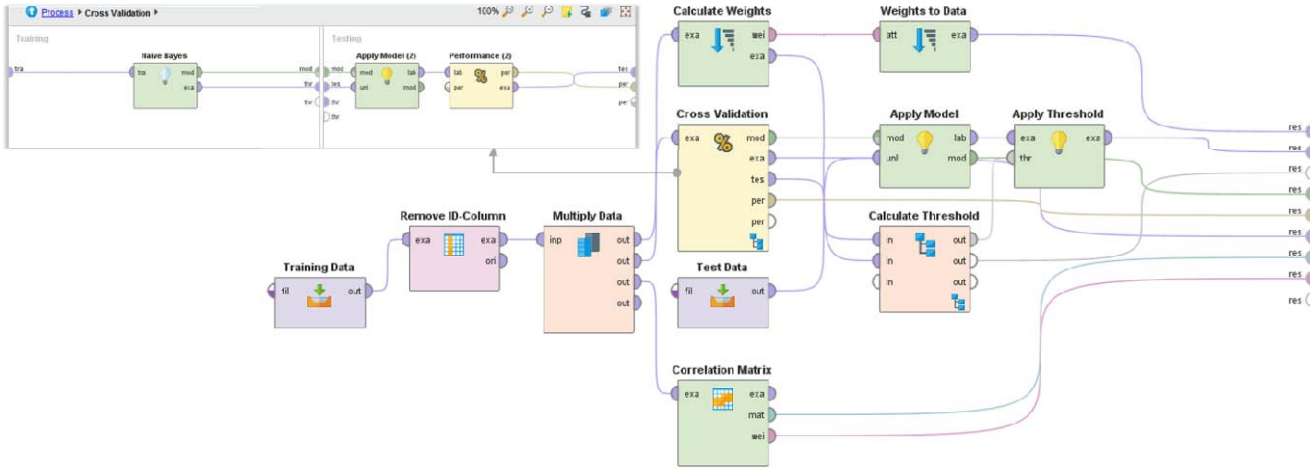

Fig. 2 Process of bus delay prediction model (RapidMiner Studio)

TABLE II
TEST RESULTS FOCUSING ON THE CORRELATION BETWEEN WEATHER AND ROAD CONDITION

Test	Training data (precipitation (mm))	Test data (precipitation (mm))	Data ratio	Prediction result (%)	Accuracy
1	Sep. 22 (7 to 9 am/ 6 to 8 pm) (0)	Sep. 19 (7 to 9 am/ 6 to 8 pm) (7.8)	8 : 2	Correct: 90.67 Fail: 9.33	96.58% +/- 1.42%
2	Sep. 22 (7 to 9 am/ 6 to 8 pm) (0)	Sep. 19 (7 to 9 am/ 6 to 8 pm) (7.8)	7 : 3	Correct: 88.61 Fail: 11.39	95.71% +/- 1.61%
3	Sep. 25 (7 to 9 am) (0)	Sep. 27 (7 to 9 am) (1.3)	1 : 1	Correct: 63.55 Fail: 36.45	96.15% +/- 1.87%
4	Sep. 19 (7 to 9 am (0)/ 6 to 8 pm (7.8))	Sep. 27 (8 to 9 am) (1.3)	8 : 2	Correct: 90.99 Fail: 9.01	96.14% +/- 1.33%
5	Sep. 20 to 22 (7 to 9 am) (0)	Sep. 25 (8 to 9 am) (0)	8 : 2	Correct: 91.37 Fail: 8.63	97.21% +/- 0.62%
6	Sep. 20 to 22 (7 to 9 am) (0) + Sep. 25 (7 to 8 am) (0)	Sep. 25 (8 to 9 am) (0)	8 : 2	Correct: 96.27 Fail: 3.73	98.02% +/- 0.76%

C. Application of the Prediction Model and Test

The prediction model is tested several times as classifying types. The classification types are separated by affecting the weather, containing data of test day, and accumulating amount of data. Each test judges the accuracy of prediction by comparing real operating status.

The weather condition is a close correlation with the road status [13]. The test is conducted under the assumptions that rainy day effects road congestion, so it causes a problem of managing bus headway. Test 1 tries to predict rain day headway issue with the data of clear day. Test 2 and test 3 is a verification of the Test 1, how the data amount and data ratio will give effects for accuracy prediction. Test 4 examines the difference of prediction accuracy between the test 1 and the prediction of rainy day by another rainy day. The prediction results of each test are illustrated in Table 2.

The accumulated data are continuous by a minute interval, and the data has a series of continuous bus stops. The amount of training data and test data are difficult to be same amount in every experiment. Therefore, tests are conducted with identical data ratio, not the number of data.

The experiments, test1 and test 2, show that high ratio of training data gains high prediction result. However, the prediction result is rather lower than other tests as expected even though using a large number of data. This is because the ratio between training and test data is inappropriate. On the contrary, this result means that the prediction accuracy could be increased by plenty of data, even though the tests have same conditions exactly.

In the Test 4, it has precipitation data both of training and test data. And it is predicted by accumulating data on different days. The test has different raining time on training and test data. Although both the two data sets have precipitation as attribute, test 4 is not appeared clear differences by comparing the predicted probability of test 1 because time acts as attribute. The impact is low since it has the difference of time. However, the prediction accuracy on rainy day could be increased by accumulating data on rainy day.

Test 5 and test 6 examines the impact of the prediction accuracy by implementing real-time data into the previous test rule. Test 6 has basically same test condition with Test 5, but it has used more training data, one-hour earlier bus delay information. This test experiments the difference between the training with past days data only and the training with past days data and target days data additionally.

V. PREDICTION RESULTS ON BUS DELAY

The tests focusing on precipitation draws the results as follows. The prediction accuracy of test 1 is similar to test 4. However, the weight values of training data set have differences as Table 3. The main difference is precipitation between test 1 and test 4. It appears in weight values. Test 4 presents precipitation (W-precipitation), precipitation type (W-type), and road speed (T-speed) as important factors than test 1. Both data sets examine delay status focusing on intervals. As result, bus interval (B-interval) is appeared as the most important factor in common. And the second factor is prediction time of the second bus (B-predictT2).

The prediction results of test 5 (Fig. 3) and test 6 (Fig. 4) as follows. 'Prediction (IntervalType2)' is prediction result about the label by learning model. This represents with each confidence about '0 (delay)' and '1 (No)'. The label is judged by the confidence. The wrong predictions of test 5 are the 44 examples among the 510 examples. Test 6 has the 19 wrong prediction examples.

TABLE III
COMPARISON OF THE WEIGHT VALUE BETWEEN TEST 1 AND TEST 4

Attribute	Weight value of test 1	Weight value of test 4
B-interval	1	1
B-predictT2	0.361	0.334
examples-time	0.032	0.082
W-temp	0.026	0.074
W-precipitation	0	0.033
W-type	0	0.012
T-speed	0.005	0.007
B-predictT1	0.016	0.007
B-stop	0.001	0

Fig. 5 presents the relation of 'IntervalType2', 'Prediction (IntervalType2)', and 'B-interval' as scatter graphs based on the result of test 5 and test 6. Test 5 doesn't have examples classifying as '0'. This is not classified properly because the

data set of Sep. 20 to 22 (2397 examples) doesn't have matching examples to the test case. When the amount of training data set extends to Sep. 18 to 22 (6500 examples), the wrong predictions are deduced 4 examples among the 510 test examples. Fig. 6 represents that the 'correct' prediction is possible by accumulating the amount of data.

ExampleSet (510 examples, 4 special attributes, 7 regular attributes) Filter (44 / 510 examples): wrong_predictions

Row No.	IntervalType2	Prediction(IntervalType2)	confidence(1)	confidence(0)	examples-time	W-temp	T-speed	B-stop	B-predictT1	B-predictT2	B-interval
1	0	1	1.000	0.000	Sep 25, 2017 8:12:05 AM KST	18800	18	217000086	5	18	13
2	0	1	1.000	0.000	Sep 25, 2017 8:18:30 AM KST	18800	13	217000315	1	13	12
3	0	1	1.000	0.000	Sep 25, 2017 8:36:13 AM KST	18800	13	217000315	1	13	12
4	0	1	1.000	0.000	Sep 25, 2017 8:18:31 AM KST	18800	20	217000314	2	14	12
5	0	1	1.000	0.000	Sep 25, 2017 8:38:35 AM KST	18800	14	217000314	1	12	11
6	0	1	1.000	0.000	Sep 25, 2017 8:18:31 AM KST	18800	5	216000118	4	15	11
7	0	1	1.000	0.000	Sep 25, 2017 8:38:35 AM KST	18800	6	216000118	2	13	11
8	0	1	1.000	0.000	Sep 25, 2017 8:18:31 AM KST	18800	12	216000117	6	17	11
9	0	1	1.000	0.000	Sep 25, 2017 8:42:18 AM KST	18800	17	216000074	1	12	11
10	0	1	1.000	0.000	Sep 25, 2017 8:44:03 AM KST	18800	2	216000389	3	14	11
11	0	1	1.000	0.000	Sep 25, 2017 8:46:21 AM KST	18800	2	216000389	1	13	12
12	0	1	1.000	0.000	Sep 25, 2017 8:08:05 AM KST	18800	19	216000073	1	12	11
13	0	1	1.000	0.000	Sep 25, 2017 8:46:21 AM KST	18800	34	216000073	2	14	12
14	0	1	1.000	0.000	Sep 25, 2017 8:08:05 AM KST	18800	36	217000567	2	13	11
15	0	1	1.000	0.000	Sep 25, 2017 8:46:21 AM KST	18800	40	217000567	2	14	12
16	0	1	1.000	0.000	Sep 25, 2017 8:08:05 AM KST	18800	14	217000264	3	14	11
17	0	1	1.000	0.000	Sep 25, 2017 8:48:07 AM KST	18800	7	217000264	2	13	11
18	0	1	1.000	0.000	Sep 25, 2017 8:30:15 AM KST	18800	19	217000263	1	12	11
19	0	1	1.000	0.000	Sep 25, 2017 8:08:03 AM KST	18800	15	217000271	7	18	11
20	0	1	1.000	0.000	Sep 25, 2017 8:14:36 AM KST	18800	15	217000271	1	13	12
21	0	1	1.000	0.000	Sep 25, 2017 8:30:15 AM KST	18800	18	217000271	2	13	11
22	0	1	1.000	0.000	Sep 25, 2017 8:32:19 AM KST	18800	17	217000271	1	12	11
23	0	1	1.000	0.000	Sep 25, 2017 8:50:11 AM KST	18800	10	217000271	3	14	11
24	0	1	1.000	0.000	Sep 25, 2017 8:52:05 AM KST	18800	24	217000271	1	13	12
25	0	1	1.000	0.000	Sep 25, 2017 8:14:37 AM KST	18800	30	217000270	1	14	13
26	0	1	1.000	0.000	Sep 25, 2017 8:30:15 AM KST	18800	34	217000270	3	14	11
27	0	1	1.000	0.000	Sep 25, 2017 8:32:19 AM KST	18800	33	217000270	1	13	12
28	0	1	1.000	0.000	Sep 25, 2017 8:52:05 AM KST	18800	13	217000270	2	14	12
29	0	1	1.000	0.000	Sep 25, 2017 8:14:37 AM KST	18800	19	217000337	3	15	12
30	0	1	1.000	0.000	Sep 25, 2017 8:16:15 AM KST	18800	23	217000337	1	13	12
31	0	1	1.000	0.000	Sep 25, 2017 8:32:19 AM KST	18800	39	217000337	1	14	13
32	0	1	1.000	0.000	Sep 25, 2017 8:14:37 AM KST	18800	6	216000330	5	17	12
33	0	1	0.999	0.001	Sep 25, 2017 8:16:15 AM KST	18800	6	216000330	2	16	14
34	0	1	1.000	0.000	Sep 25, 2017 8:18:30 AM KST	18800	6	216000330	2	13	11
35	0	1	1.000	0.000	Sep 25, 2017 8:34:04 AM KST	18800	6	216000330	3	14	11
36	0	1	1.000	0.000	Sep 25, 2017 8:56:06 AM KST	18800	8	216000330	1	13	12
37	0	1	1.000	0.000	Sep 25, 2017 8:14:37 AM KST	18800	28	216000367	6	18	12
38	0	1	1.000	0.000	Sep 25, 2017 8:16:15 AM KST	18800	28	216000367	3	16	13
39	0	1	1.000	0.000	Sep 25, 2017 8:18:31 AM KST	18800	28	216000367	3	14	11
40	0	1	1.000	0.000	Sep 25, 2017 8:20:23 AM KST	18800	28	216000367	1	12	11
41	0	1	1.000	0.000	Sep 25, 2017 8:04:09 AM KST	18800	5	216000070	1	12	11
42	0	1	1.000	0.000	Sep 25, 2017 8:16:15 AM KST	18800	6	216000070	5	18	13
43	0	1	1.000	0.000	Sep 25, 2017 8:18:31 AM KST	18800	6	216000070	4	16	12
44	0	1	1.000	0.000	Sep 25, 2017 8:20:24 AM KST	18800	6	216000070	3	14	11

Fig. 3 The prediction result of test 5

Both tests similar in the weight value but the priority of 'W-temp' and 'example-time' is different as shown in Table 4.

Test 5 appears the effect of temperature clearly more than test 6. Test 6 indicates the influence of 'example-time' than test 5. This result is considered as influencing the sample day data sets to the final prediction result. Also, test 6 containing test day data is high in the final prediction accuracy. As a result, data set containing test day in real-time is useful to predict bus circumstance. Fig. 7 compares the result of test 5 and test 6 through the correlation matrix.

VI. Discussion

Implementing a big data and machine learning technology to the city science is inevitable. The city produces a huge volume of data every second. The data is traces of citizens' life and measures for livability. Cities are the one of the most complex system that require such technologies which can give us insight of the complexity, by weighting of relations among diverse urban events.

ExampleSet (510 examples, 4 special attributes, 7 regular attributes) Filter (19 / 510 examples): wrong_predictions

Row No.	IntervalType2	Prediction(IntervalType2)	confidence(1)	confidence(0)	examples-time	W-temp	T-speed	B-stop	B-predictT1	B-predictT2	B-interval
1	0	1	0.897	0.103	Sep 25, 2017 8:38:35 AM KST	18.800	14	217000314	1	12	11
2	0	1	0.854	0.146	Sep 25, 2017 8:38:35 AM KST	18.800	6	216000118	2	13	11
3	0	1	0.900	0.100	Sep 25, 2017 8:42:18 AM KST	18.800	17	216000074	1	12	11
4	0	1	0.805	0.195	Sep 25, 2017 8:44:03 AM KST	18.800	2	216000389	3	14	11
5	0	1	0.901	0.099	Sep 25, 2017 8:08:05 AM KST	18.800	19	216000073	1	12	11
6	0	1	0.877	0.123	Sep 25, 2017 8:08:05 AM KST	18.800	36	217000567	2	13	11
7	0	1	0.797	0.203	Sep 25, 2017 8:08:05 AM KST	18.800	14	217000264	3	14	11
8	0	1	0.851	0.149	Sep 25, 2017 8:48:07 AM KST	18.800	7	217000264	2	13	11
9	0	1	0.899	0.101	Sep 25, 2017 8:30:15 AM KST	18.800	19	217000263	1	12	11
10	0	1	0.853	0.147	Sep 25, 2017 8:30:15 AM KST	18.800	18	217000271	2	13	11
11	0	1	0.898	0.102	Sep 25, 2017 8:32:19 AM KST	18.800	17	217000271	1	12	11
12	0	1	0.800	0.200	Sep 25, 2017 8:50:11 AM KST	18.800	10	217000271	3	14	11
13	0	1	0.828	0.172	Sep 25, 2017 8:30:15 AM KST	18.800	34	217000270	3	14	11
14	0	1	0.853	0.147	Sep 25, 2017 8:18:30 AM KST	18.800	6	216000330	2	13	11
15	0	1	0.802	0.198	Sep 25, 2017 8:34:04 AM KST	18.800	6	216000330	3	14	11
16	0	1	0.818	0.182	Sep 25, 2017 8:18:31 AM KST	18.800	28	216000367	3	14	11
17	0	1	0.908	0.092	Sep 25, 2017 8:20:23 AM KST	18.800	28	216000367	1	12	11
18	0	1	0.899	0.101	Sep 25, 2017 8:04:09 AM KST	18.800	5	216000070	1	12	11
19	0	1	0.901	0.099	Sep 25, 2017 8:20:24 AM KST	18.800	6	216000070	3	14	11

Fig. 4 The prediction result of test 6

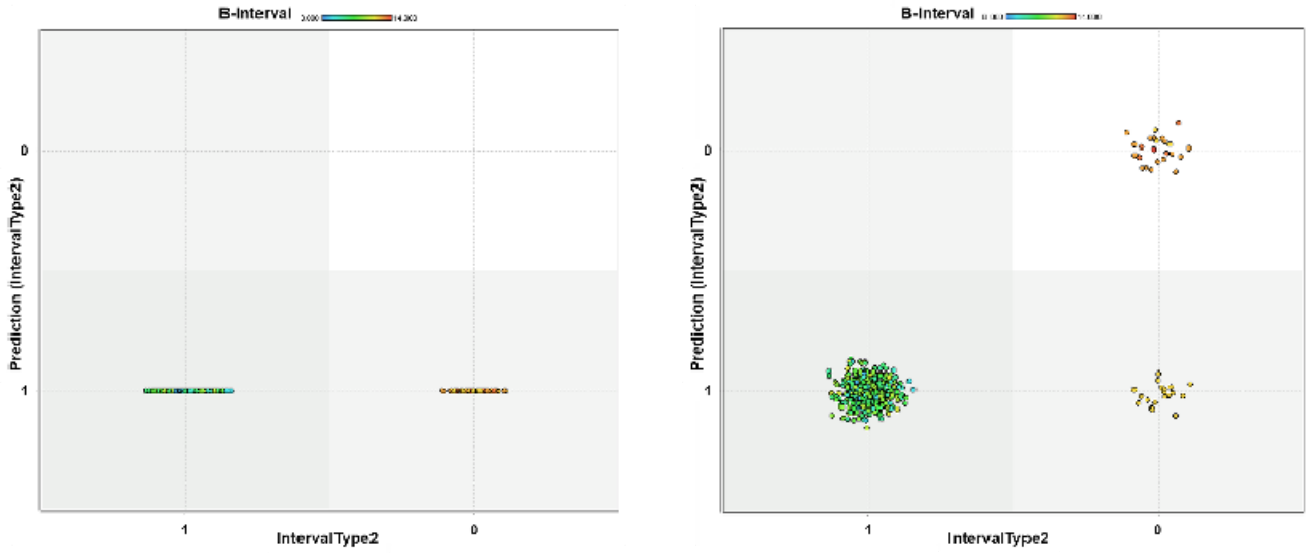

Fig. 5 Scatter graphs about the relations of IntervalType2, Prediction, and B-interval of test 5 and test 6

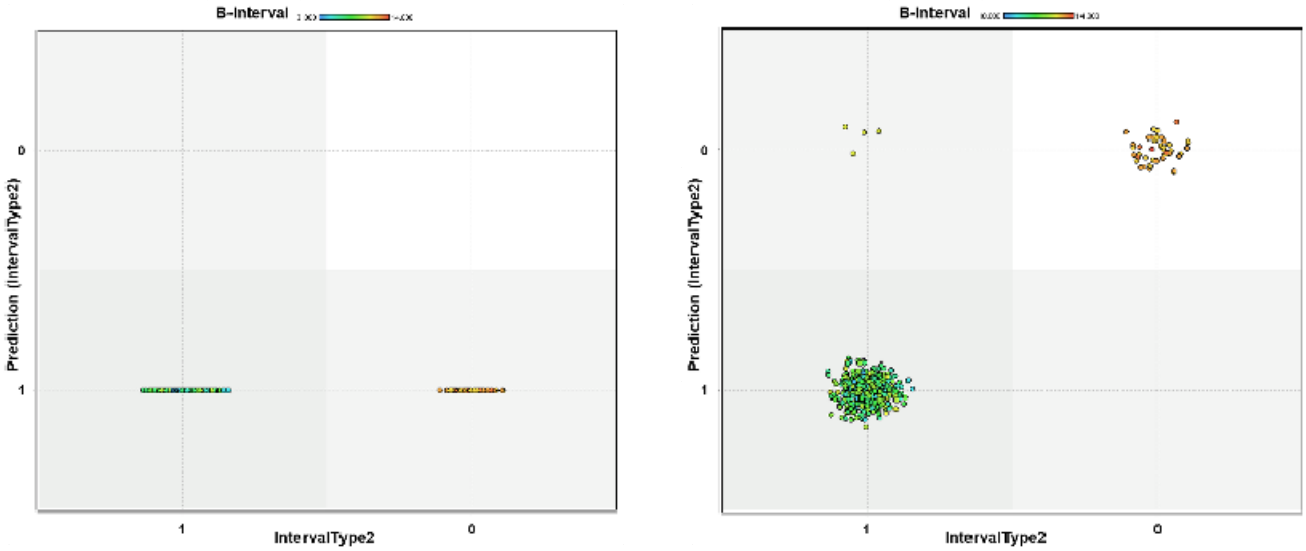

Fig. 6 Changes of the prediction result according to the amount of data of test 5 (the relations of IntervalType2, Prediction, and B-interval)

Attributes	examples-time	W-temp	T-speed	B-stop	B-predictT1	B-predictT2	B-interval
examples-time	1	0.043	0.008	-0.000	0.062	-0.036	-0.109
W-temp	0.043	1	0.022	-0.000	-0.141	-0.077	0.035
T-speed	0.008	0.022	1	0.352	-0.023	-0.030	-0.018
B-stop	-0.000	-0.000	0.352	1	0.003	0.006	0.005
B-predictT1	0.062	-0.141	-0.023	0.003	1	0.666	-0.088
B-predictT2	-0.036	-0.077	-0.030	0.006	0.666	1	0.684
B-interval	-0.109	0.035	-0.018	0.005	-0.088	0.684	1

Attributes	examples-time	W-temp	T-speed	B-stop	B-predictT1	B-predictT2	B-interval
examples-time	1	-0.870	0.054	-0.000	-0.009	0.026	0.042
W-temp	-0.870	1	-0.051	0.000	0.004	-0.061	-0.084
T-speed	0.054	-0.051	1	0.346	-0.023	-0.023	-0.009
B-stop	-0.000	0.000	0.346	1	0.001	0.004	0.004
B-predictT1	-0.009	0.004	-0.023	0.001	1	0.649	-0.110
B-predictT2	0.026	-0.061	-0.023	0.004	0.649	1	0.685
B-interval	0.042	-0.084	-0.009	0.004	-0.110	0.685	1

Fig. 7. Correlation matrix of test 5 and test 6

TABLE IV
COMPARISON OF THE WEIGHT VALUE BETWEEN TEST 5 AND TEST 6

Attribute	Weight value of test 5	Weight value of test 6
B-interval	1	1
B-predictT2	0.285	0.269
examples-time	0.047	0.044
W-temp	0.049	0.033
W-precipitation	0.022	0.022
W-type	0.002	0.002
T-speed	0	0
B-predictT1	0	0
B-stop	0	0

The public transportation is a difficult subject, since it has complicated invisible relations with other environmental aspects. In order to make the invisible visible, this research conducts an experiment, bus delay prediction, by using urban big data and machine learning technology.

This research focused on predicting irregular bus event happenings, and the result shows high accuracy in the prediction. It would contribute system enhancement by real-time based bus interval control system. This research is conducted in the prototyping level by focusing on methodology considering possibility of the weak A.I technology. If the public institutions utilize this methodology, the methodology is able to apply to construct the intelligence transport system (ITS).

The ITS is the transportation system to improve the

efficiency and safety of the traffic and to make the automation and scientification for operating and managing the system by utilizing high technologies and information [14], [18].

The ITS functions to the whole city as a target to enhance traffic communications or operate traffic signals for responding to real-time according to the traffic conditions. Furthermore, autonomous vehicle system is expected to apply to the city in near future. The USA and some of European countries such as Switzerland and Germany are conducting the trial tests of autonomous vehicle in real environment. Also, Korea is planning the trial tests in December 2017. As a result, achievements of this research should be utilized helpfully as a methodology according to the introduction of these intelligence systems.

ACKNOWLEDGMENT

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (NRF-2016R1C1B2013424).

REFERENCES

[1] Korea Planners Association, *Urban Planning*, Bosunggak, 2009, pp. 36-37.
[2] J. Gubbi, R. Buyya, S. Marusic and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future generation computer systems*, vol. 29, no. 7, 2013, pp. 1645-1660.
[3] W.-K. Lee, M.-K. Kim, Y.-S. Kim and J.-H. Lee, "Study on implementation plan of flexible headway service of city bus," Busan Development Institute, 2009. (in Korea)
[4] K.-W. Kim, "Study on the city bus use demand and flexible service during precipitation," Ph. D. Dissertation, Pusan National University, 2012. (in Korea)
[5] S.-J. Lee, "Big Data for Transportation Policies and Their Applications," *The Korea Transport Institute*, 2013. (in Korea)
[6] J.-W. Yi and I.-K. Kim, "A Study on The Integrate Evaluation of Urban Bus Service in Seoul," *Journal of Transport Research*, vol. 20, no. 4, 2013, pp. 131-145. (in Korea)
[7] L. Eboli and G. Mazzulla, "A Methodology for Evaluating Transit Service Quality Based on Subjective and Objective Measures from the Passenger's Point of View," *Transport Policy*, vol. 18, issue 1, 2011, pp. 172-181.
[8] M. Friman, "Implementing Quality Improvements in Public Transport," *Journal of Public Transportation*, vol. 7, no. 4, 2004, pp. 49-65.
[9] T. Liebig, N. Piatkowski, C. Bockermann and K. Morik, "Dynamic route planning with real-time traffic predictions," *Information Systems*, vol. 64, 2017, pp. 258-265.
[10] H. S. Lee, J. H. Park, S. H. Jo and B. J. Yun, "Development of Optimal Bus Scheduling Algorithm with Multi-constraints," *Journal of Korean Society of Transportation*, vol. 24, no. 7, 2006, pp. 129-138. (in Korea)
[11] M. Ruan and J. Lin, "An investigation of bus headway regularity and service performance in Chicago bus transit system," *in Transport Chicago, Annu. Conf.*, Vol. 14, June 2009.
[12] S.-Y. Ko, J.-S. Ko and J.-S. Jeon, "Development of Real Time Vehicle Scheduling Model for Public Transportation," *Journal of the Research Institute of Industrial Technology*, vol. 18, 1999, pp. 181-186 (in Korean)
[13] T. Maze, M. Agarwai and G. Burchett, "Whether weather matters to traffic demand, traffic safety, and traffic operations and flow," *Transportation research record: Journal of the transportation research board*, no. 1948, 2006, pp. 170-176.
[14] C. Dobre and F. Xhafa, "Intelligent services for big data science", *Future Generation Computer Systems*, vol. 37, 2014, pp. 267-281.
[15] Open Data Portal - https://www.data.go.kr/ (2017. 10. 9.)
[16] Gyeonggi Bus Information System (GBIS) - http://www.gbis.go.kr/ (2017. 10. 9.)
[17] National Weather Center - https://data.kma.go.kr/cmmn/main.do (2017. 10. 9.)
[18] Intelligent Transport Society of Korea - http://www.itskorea.kr/02_sta/sta1.jsp (2017. 10. 9.)