Real-Time Travel Time Prediction Using Multi-level k-Nearest Neighbor Algorithm and Data Fusion Method

Sehyun Tak[1], Sunghoon Kim[2], Kiate Jang[3] and Hwasoo Yeo[4]

[1]Smart Transportation System Laboratory, Department of Civil and Environmental Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea; PH (42) 350-3674; email: taksehyun@kaist.ac.kr
[2]Smart Transportation System Laboratory, Department of Civil and Environmental Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea; PH (42) 350-3674; email: sunghoon.kim@kaist.ac.kr
[3]Graduate School for Green Transportation, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea; PH (42) 350-1264; email: kitae.jang@kaist.ac.kr
[4]Smart Transportation System Laboratory, Department of Civil and Environmental Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea; PH (42) 350-3634; email: hwasoo@kaist.edu

ABSTRACT

Estimating and predicting the travel time on freeways with reasonable accuracy is essential for successful implementation of intelligent transportation system. However, the related previous studies have some problems. The statistics-based methods have problems in accuracy, and some others are limited to predicting the travel time only during short time interval. Another challenging matter is that the existing road sensors have some limitations in being directly utilized, because data from road sensors have lots of errors. In this study we propose a new algorithm called multi-level k-Nearest Neighbor (k-NN), which is designed for predicting travel time with higher computational efficiency and prediction accuracy. The algorithm consists of three parts: classification, global matching, and local matching. As a part of the proposed algorithm, in order to overcome the problems of data errors, we provide a data fusion method that combines the traffic data from ILDs and DSRC. The results show that the proposed multi-level k-NN with the data fusion can effectively predict the future travel time within less than 5% error range, even in congested traffic situations.

INTRODUCTION

Utilizing the existing transportation infrastructure more efficiently and intelligently, rather than just extending it, is becoming more important for managing road traffic management and providing traveler information. Intelligent Transportation Systems are intended to increase the road use efficiency and to provide a high level of automation in freeway systems, by using advanced traveler information systems. Estimating and predicting freeway travel time with reasonable accuracy are the ones that are important for the successful implementation of

Intelligent Transportation Systems, because it can provide useful information to road system users. So, many studies have been conducted and several number of travel time prediction models have been developed using the information gathered from various sources such as Inductive Loop Detectors (ILDs), Dedicated Short Range Communications (DSRC), Toll Collection System (TCS), and probe vehicles. Some previous studies proposed statistical models. Gault and Taylor (1981) developed a regression model to estimate link travel time using ILDs. Rilett and Aerde (1991) predicted link travel time based on the relation between historical link travel time profiles and real-time link travel time profiles. Such work was done under the assumption that the ratio of the current travel time to the mean travel time from the historical data remains unchanged over the future time period. Indeed, these statistical models can effectively predict the future travel time in normal situations. However, the accuracy of the statistical models decreases when the real-time link travel time deviates from the mean value of the historical travel times, particularly in congested situations.

A number of researchers have examined neural networks for forecasting link travel time. Florio and Mussone (1996) applied a neural network model to predict the traffic variables such as density, flow and speed on freeways, and Park and Rilett (1999) developed a neural network model for forecasting link travel times. Cherrett et al. (2001) proposed the Artificial Neural Network (ANN) model to estimate the vehicle speed using ILDs. Neural network models predict the one or two link travel time during 5-10 minute time interval. This prediction time interval is too short to calculate the long distance travel time of the journey in the freeway. For example, in Korean expressway system, at least 3-hour travel time prediction capability is required to provide the travel time information to the road system users. There are also other studies like Takahashi et al. (1996), which predict the travel time based on the Kalman filtering model. These studies, including the neural network based models, are limited to predicting the travel time only during short time interval. Furthermore, they practice the matter under only one or two types of traffic conditions, such as free flow and congestion.

Another challenging matter in predicting travel time is that the existing road sensors have many errors, so, there are some limitations in fully utilizing them. In this context, fusing the freeway traffic data such as spot speeds, in/out traffic flow, and link travel time collected from ILDs, Toll Collection System (TCS), and DSRC is the critical task. This task is challenging, because each data differs in terms of spatial resolution, accuracy, detector health, and calculation method.

The objective of this study is to develop a new model that can efficiently predict travel times in long-term, by using real-time and historical traffic data. The new algorithm to be proposed is called multi-level k-Nearest Neighbor (k-NN). Such algorithm consists of three consecutive processes for traffic pattern matching and predicting future travel time with higher accuracy and lower computation time. As a part of the algorithm, we provide a data fusion method that combines the traffic data from ILDs and DSRC. The purpose of practicing this method is to reduce the data error, so that it can later raise the accuracy of the prediction results, as it gets to be applied to the algorithm. We use the term, hybrid data, for this combined data.

MULTI-LEVEL K-NEAREST NEIGHBOR ALGORITHM

We propose the multi-level k-Nearest Neighbor algorithm for predicting long-term travel time on freeways based on the k-NN classification rule. The k-NN algorithm offers a simple and accurate way of undertaking classification. Figure 1 shows the entire framework of the proposed multi-level k-NN method. As shown in Figure 1, the proposed method consists of three different layers: classification, global matching, and local matching. The descriptions of each layer are as follows.

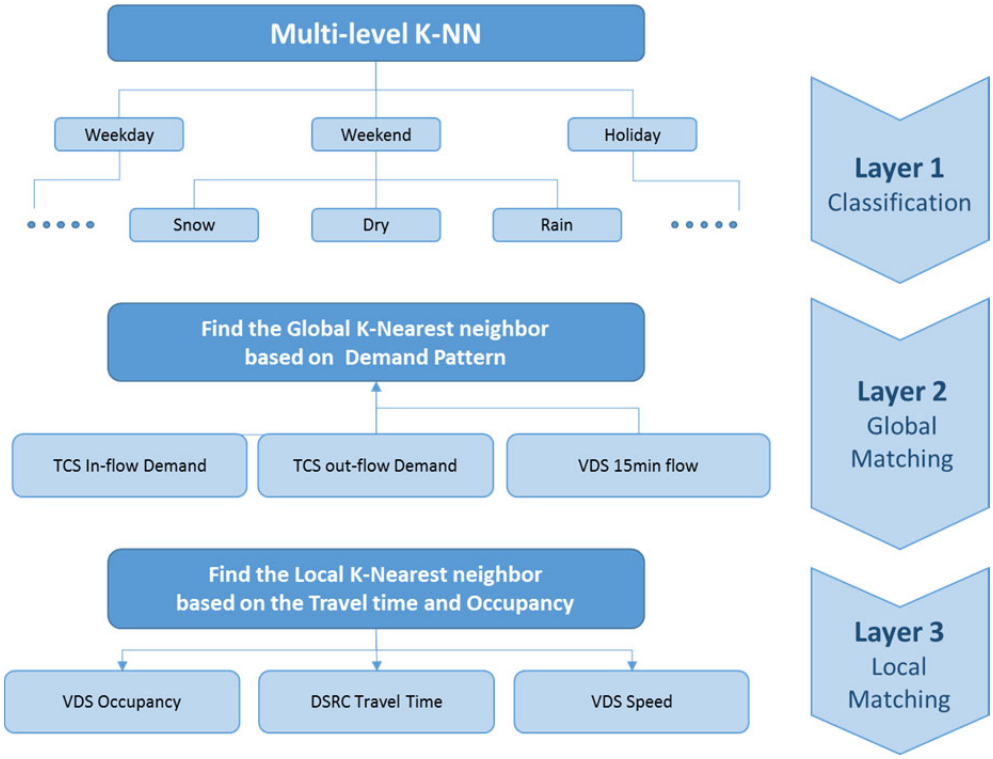

Figure 1. The framework of multi-level k-NN algorithm

Layer 1. Classification. In this layer of classification, the day types (weekdays, weekend, and holidays) and weather types are classified before actually applying the k-NN pattern matching. Traffic demand and travel time patterns are quite different depending on the day and weather types. Therefore, to reduce the searching space of historical data for pattern matching for lowering the computation time of the entire algorithm, we subdivide the dates based on the day type and weather type.

Layer 2. Global Matching. In the layer of global matching, the algorithm finds the most 30 similar days in terms of traffic demand pattern, by using TCS data (1 hour interval) and ILDs (15 minutes interval). Figure 2 shows the global matching method in the multi-level k-NN algorithm. As shown in Figure 2, the inflow and outflow of a global section are estimated based on the information from the loop detectors. The demand is estimated based on the information from the tollgates located inside the global section. By using these two types of data, the Euclidean distances of flow and

demand patterns are calculated. The Euclidean distance represents the difference between the real-time data and historical data. The less Euclidean distance represents the historical day that has more similar flow and demand patterns.

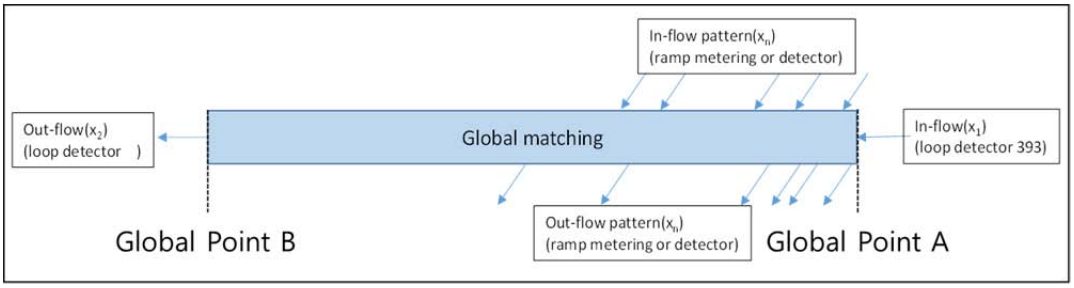

Figure 2. The illustration of the global matching

Layer 3. Local Matching. In the layer of local matching, based on the global matching results, the local matching is conducted. Global section consists of several consecutive local sections as shown in Figure 3. In this layer, multi-level k-NN algorithm finds the most 5 similar days in terms of occupancy and speed patterns from ILDs (5 minute interval) and DSRC (5 minutes interval) data, respectively. In the local matching layer, the Euclidean distance of each local section is calculated to find the historical days having the similar speed and occupancy patterns in the 30 day that was previously resulted from the global matching process. The future travel time is predicted by using the historical data based on the assumption that the date having the most similar speed pattern from certain time (e.g. 4 hours ago) until the current time (the real-time) will indicate the most similar speed pattern in the future.

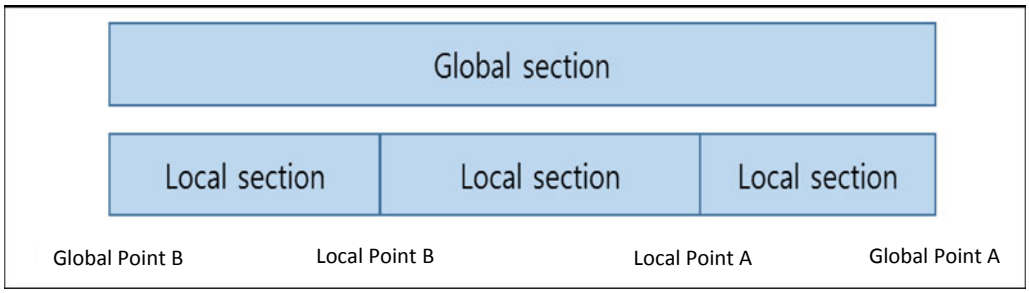

Figure 3. A brief illustration of global and local matching sections

Data Fusion Method. As mentioned earlier, predicting travel time with a single data source may have a problem in accuracy, because the existing data sources have many errors. The errors come from maintenance issue, weather conditions, and etc. When we use the historical patterns for matching the k-NN, such process would include missing data points. If there are many errors in both the real-time and historical data, it would result as they have the similar traffic patterns, even though the traffic situations of the two different days differ from each other. For example, in the worst case, if all real-time data points are missing and if all historical data points from a road section are also missing, it would result as the perfect match. This would result some serious errors in predicting the future travel time.

So, to reduce such effect, we combined two heterogeneous speed data from each ILDs and DSRC. As shown in Figure 4, when the speed value of DSRC data is missing, we can instead use the speed value of ILDs to fill out the information on the missing data point. On the other hand, the speed value of DSRC data can cover the missing data point in ILDs speed values. Them, it will result the speed values of an entire section with the least amount of data missing, thus, the speed data will be at the minimum level of errors. We call this speed, which is the result of the data fusion method, as the "hybrid speed." By using the hybrid speed, not only the accuracy of travel time prediction would be increased, but also the reliability of the traffic data itself would be enhanced.

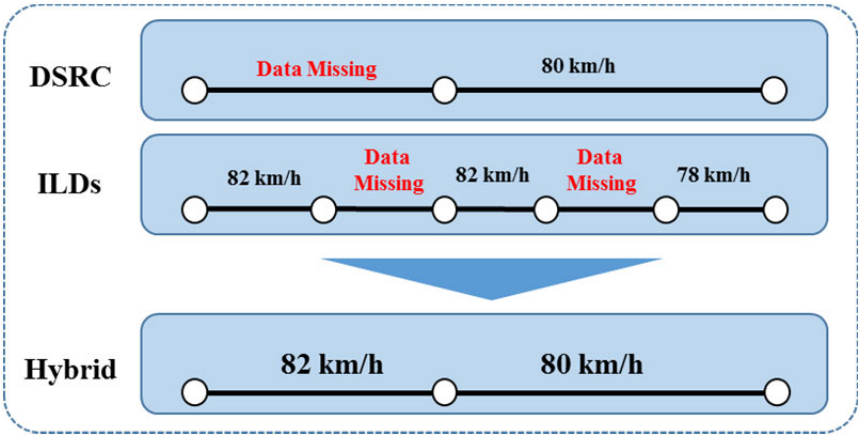

Figure 4. Data fusion methodology for reducing data errors

TRAVEL TIME PREDICTION RESULT

The data we used for this study is derived from the Kyungbu expressway system (Highway No. 1) in Republic of Korea. The measurement used in this study is made between Daejeon and Seoul city. In this section, the road has four or five lanes in each direction. The lanes are numbered consecutively from 1 at the offside to 5 at the nearside. The data used in the present analysis represents all available records from the 1st day of January to the 31st day of August in 2013.

Figure 5 shows the prediction results of the proposed multi-level k-NN algorithm. The figure shows the travel time prediction result in the congested traffic condition. It shows that the proposed algorithm can accurately predict the future travel time. In this site the maximum delay appears near 10:00 am and the travel time increases by approximately 60 minutes compared to the free flow traffic condition. At the most congested time period, the proposed algorithm can predict the travel time with error of less than 5 minutes.

However, the error of approximately 15 minutes occurred during the transition time from free flow to congestion state, and vice versa. This type of error occurred because k-NN methodology needs certain traffic characteristics in order to find a similar travel time pattern. So, the accuracy would be guaranteed after traffic characteristic pattern begin to appear.

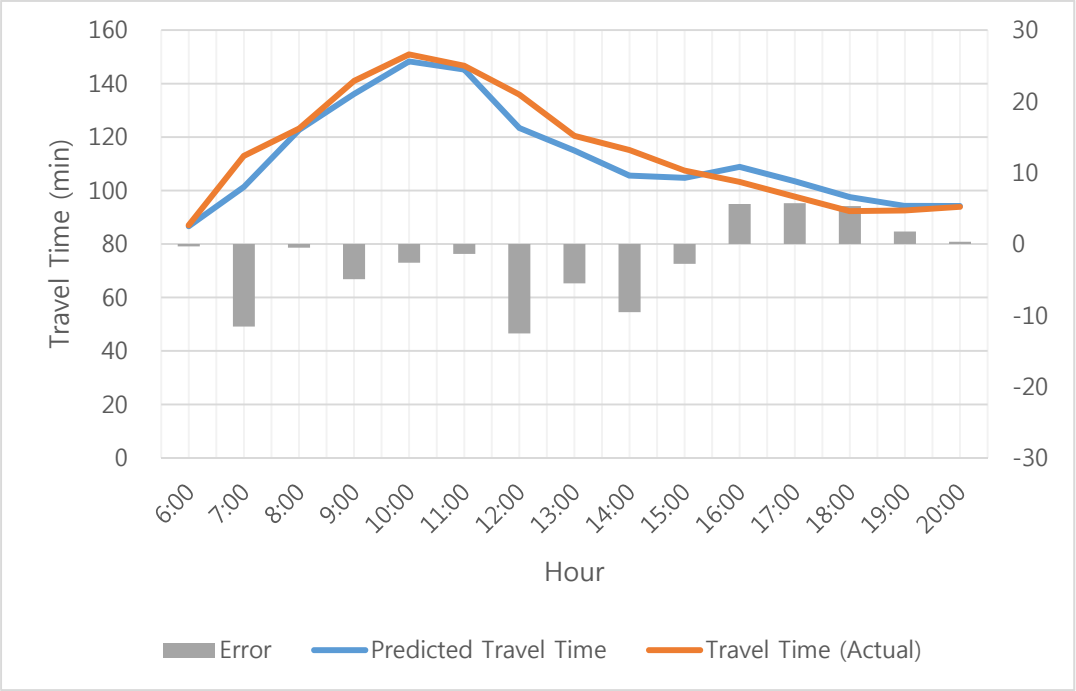

Figure 5. Travel time prediction result in a congested traffic condition

The main point of this study is to see the effect of separating the pattern recognition processes into two different layers: the global matching and local matching. "Global Matching Only" stands for the traditional k-NN algorithm, which does not have multi layers of recognizing the traffic patterns. Using the same data sources from the same road section, we predicted the travel time in two different ways. We predicted the travel time based only on the global matching results, and we also predicted the travel time based on both the global and local matching results.

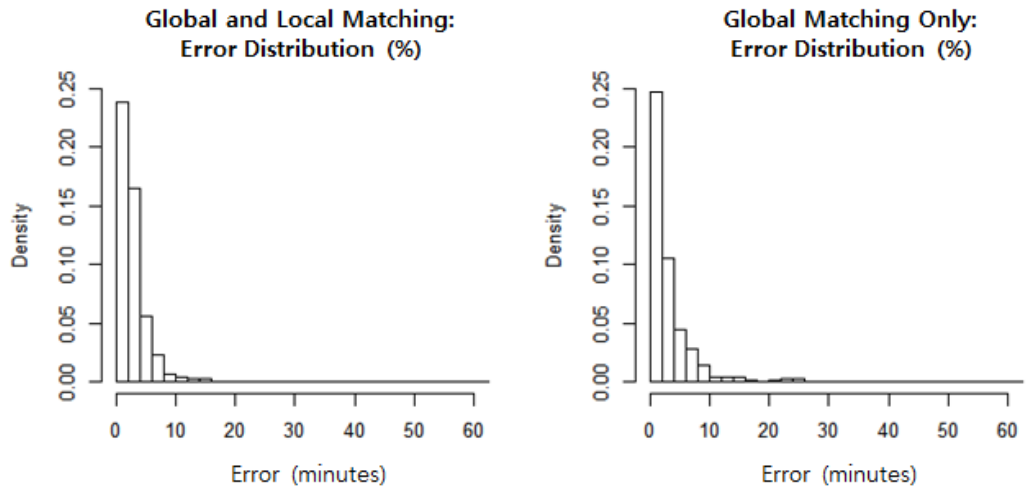

Figure 6. Error comparison between the two different measurement methods

Figure 6 shows the comparison of the error distributions of the two different prediction results, which was derived from predicting the travel time for seven consecutive days. In figure 6, it could be seemed as the Global Matching Only has the better efficiency, since the occurrences of the error less than 2 minutes are more frequent. However, we have to put our focus more on the horizontal view. As we look at the graphs in horizontal view, the percentage of the big error (more than 10 minutes) occurrences in the Global Matching Only is greater. The maximum time difference is 26 minutes for the Global Matching Only, whereas the maximum difference is 16 minutes for the "Global and Local Matching" case. It means that this type of prediction method would more likely result greater time differences, thus the error would be greater. This result shows that using the multi-layers of recognizing the traffic patterns can reduce the big error occurrences.

CONCLUSION

The proposed multi-level k-NN algorithm consists of three layers: classification, global matching, and local matching. As the result of predicting travel time, the proposed algorithm can predict the future travel time during long time intervals with high accuracy. The accuracy of the proposed algorithm is significantly affected by the detector health, because the algorithm finds the similar traffic pattern by comparing the real-time data and historical data, regardless of the data errors. In order to ensure the reliability of the prediction result, we also provided the data fusion method combining the speed data of ILDs and DSRC.

One of the merits of the multi-level k-NN algorithm compared to the traditional one is that it can predict long-term travel time. Also, in each layer, the searching space for pattern recognition is gradually reduced, so the computation time for the entire predicting process is reduced. Moreover, the percentage of the big error occurrences is reduced with the proposed algorithm, meaning that the prediction accuracy can be enhanced compared to the traditional k-NN algorithm.

The matter that needs to be considered in the proposed algorithm is the amount of the historical data. The algorithm particularly depends on both the quality and quantity of the historical data. In this study, the amount of historical data used was only 244 days (from Jan. 1 to Aug. 31). Considering the classification process by day and weather types, the amount of the sampled historical data for a certain day is not enough. So, due to such shortage of the sample data for traffic pattern matching, the proposed algorithm may have shown only a bit of improvement compared to the traditional one. Therefore, for further related study, we have to consider using a greater amount of historical data sources. Processing the greater amount of historical data requires the longer computation time, therefore, we should prepare other options for reducing the computation time as well.

ACKNOWLEDGMENT

This research was supported by MSIP (Ministry of Science, ICT&Future Planning), Korea, under the ITRC(Information Technology Research Center) support

program (NIPA-2013-(H0301-13-1012)) supervised by the NIPA (National IT Industry Promotion Agency)

REFERENCES

Cherrett, T., Bell, H. and McDonald, M. (2001). Estimating vehicle speed using single inductive loop detectors. *Proceedings of the Institution of Civil Engineers Transport*, vol. 147, no. 1, pp. 23-32.

Florio, L. and Mussone, L. (1996). Neural-network models for classification and forecasting of freeway traffic flow stability. Control Engineering Practice 4 (2), pp. 153-164.

Gault, H. and Taylor, I. (1981). The use of the output from vehicle detectors to assess delay in computer-controlled area traffic control systems. Transport Operations Research Group, Research Report no 31, University of Newcastle upon Tyne, U.K.

Park, D. and Rilett, L. (1999). Forecasting freeway link travel times with a multilayer feedforward neural network. Computer-Aided Civil and Infrastructure Engineering 14 (5), pp. 357-367.

Rilett, L. and Aerde, M. Van (1991). Route Based on Anticipated Travel Times. *In Proceedings of the Second International Conference on Applications of Advanced Technologies in Transportation Engineering,* ASEC, New York, pp. 183-187.

Takahashi, Y., Ikenoue, K., Yasui, K., and Kunikata, Y. (1996). Travel Time Information System (TTIS) in Mie Prefecture. *In Proceedings of the 3rd Annual World Congress on Intelligent Transportation Systems, Orlando, Florida.*

Xia, J., Huang, W. and Guo, J. (2012). A clustering approach to online freeway traffic state identification using ITS data. KSCE Journal of Civil Engineering 16 (3), pp. 426-432.