Proceedings of the 2007 IEEE
Intelligent Transportation Systems Conference
Seattle, WA, USA, Sept. 30 - Oct. 3, 2007

Prioritized Traffic Information Delivery Based on Historical Data Analysis

Hyunsung Jo, Byungwoo Lee, Yong-chan Na, Hyunjung Lee, Byonghwa Oh, Chulmin Yun, Jihoon Yang, *Member, IEEE,* Moonsoo Lee, and Minjeong Kim

***Abstract*—The main objective of this paper is to present and evaluate methods that can be used to extract valuable information from traffic data history for use in travel information and guidance systems. Through two different analyses, we identify the roads which the system should pay closer attention to. First, we introduce an approach that finds the patterns of the traffic speed which lie within the data history. Second, we discover the relations between traffic speeds and various features of roads. The result of our study let travel guidance systems know how heavy the traffic will be at any given road. We extend these approaches to tackle the problem of incident management, and designed a system that returns, the expected range of effect of an incident using road features.**

I. Introduction

In the recent years, there has been a large increase in the demand for real-time traffic information systems. Accordingly, a number of successful researches were conducted internationally [1]-[5]. However, research regarding statistical analysis on traffic data from the past – which we believe plays an essential role in building such systems – has been insufficient in Korea.

The goal of our study is to search for the patterns (tendencies of traffic which are related to the day, time, etc.) that lie within the traffic speed data history, in order to learn the present through comparison with the past. We, as well, extract the features that imply valuable information from the data history. We applied several methodologies, some inspired by the references and others newly developed, to the data which has been collected from Seoul city for three months, and were able to accomplish the specific tasks of our study.

Firstly, classic statistical techniques are applied to the domain of traffic information. We find the mean and standard deviation of the traffic speed data. Using these simple numeric values, we define an abnormality measure so the traffic information system can monitor the traffic flow at the moment.

Meanwhile, we also attempt to unveil the relation between the characteristics of the roads and their tendency to show congestion. In order to indicate the level of influence that each feature has on the traffic speed, we import the concept of mutual information.

In the later part of the paper we introduce the applications we have implemented based on our analysis on the traffic data. These applications are designed to provide traffic guidance systems with the priority of the links, so that they can inform the users accordingly. In situations where an instance has occurred, our applications will also return the expected range of the influence to the system so it can build a solution upon that information.

II. Data

As a part of the Intelligent Transport Systems project, the Ministry of Construction and Transportation, Korea has set a standard for the terminology used in the field. For the sake of convenience, we follow these standards in our study.

Nodes and ***links*** are the two basic units for real-time traffic information processing: A node is the transition point of the roads in terms of traffic speed, whereas, a link is the actual road connection that lies between two different nodes. Intersections, entrances and exits of bridges, tunnels, or overpasses are good examples of nodes. Other than the terms node and link we also borrow a set of terms which stand for various features of the road. The features and their meanings are shown in Table I [6].

The data was provided by Electronics and Telecommunications Research Institute (ETRI), Korea. At nodes that are located in the roads of Seoul, Korea, the traffic

Manuscript received April 2, 2007.

H. Jo is with Data Mining Laboratory, Department of Computer Science, Sogang University, 807 Adam Schall Hall, Sogang University, 1 Sinsu-dong, Mapo-gu, Seoul 121-742 Korea (e-mail: raptor21c@ gmail.com).

B. Lee is with Data Mining Laboratory, Department of Computer Science, Sogang University, 807 Adam Schall Hall, Sogang University, 1 Sinsu-dong, Mapo-gu, Seoul 121-742 Korea (e-mail: elva1212@ naver.com).

Y.C. Na is with Data Mining Laboratory, Department of Computer Science, Sogang University, 807 Adam Schall Hall, Sogang University, 1 Sinsu-dong, Mapo-gu, Seoul 121-742 Korea (e-mail: ycna@ sogang.ac.kr).

H. Lee is with Data Mining Laboratory, Department of Computer Science, Sogang University, 807 Adam Schall Hall, Sogang University, 1 Sinsu-dong, Mapo-gu, Seoul 121-742 Korea (e-mail: luckyhj777@ naver.com).

B. Oh is with Data Mining Laboratory, Department of Computer Science, Sogang University, 807 Adam Schall Hall, Sogang University, 1 Sinsu-dong, Mapo-gu, Seoul 121-742 Korea (e-mail: mr-five@ hanmail.net).

C. Yun is with Data Mining Laboratory, Department of Computer Science, Sogang University, 807 Adam Schall Hall, Sogang University, 1 Sinsu-dong, Mapo-gu, Seoul 121-742 Korea (e-mail: toro83@ nate.com).

J. Yang is with Data Mining Laboratory, Department of Computer Science, Sogang University, 805 Adam Schall Hall, Sogang University, 1 Sinsu-dong, Mapo-gu, Seoul 121-742 Korea (phone: +82-2-706-7931; fax: +82-2-706-7931; e-mail: yangjh@sogang.ac.kr).

M. Lee is with Telematics USC Research Division, Electronics and Telecommunications Research Institute, 161 Gajeong-dong, Yuseong-gu, Daejeon 305-700, Korea (e-mail: mslee@etri.re.kr)

M. Kim is with Telematics USC Research Division, Electronics and Telecommunications Research Institute, 161 Gajeong-dong, Yuseong-gu, Daejeon 305-700, Korea (e-mail: minjkim@etri.re.kr)

TABLE I
LINK ATTRIBUTES

Feature	Definition
LINK_ID	10-digit identification of the link
F_NODE	Starting Node ID on the link
T_NODE	Ending Node ID on the link
ROAD_USE	Passable or impassable
LANES	The number of lanes
ROAD_RANK	Road Rank consists of 7 kinds (national highway(101), local roads(103), and so on)
ROAD_TYPE	Road Type means a construction, which consists of 4 kinds (road(000), tunnel(004), and so on)
TERRITORY	The code of information on the link's location
LINK_LENGTH	The length of the link
MULTI_LINK	If link have more than two ROAD_NOs, then MULTI_LINK, else SINGLE_LINK
MAX_SPD	Permissible max speed
REST_VEH	Impassable car type which consists of 7 kinds (none(0), two-wheeled vehicle(6), and so on)
REST_W	Impassable car weight
REST_H	Impassable car height
CONNECT	CONNECT consists of 8 kinds (not connecting road(000), road connected with national highway(101), and so on)
ROAD_NO	Road number

speed has been recorded every 10 minutes from 07:00 to 11:00, 17:00 to 22:00 on weekdays and from 07:00 to 22:00 on weekends and holidays for a period of 3 months (from 11.09.2006 to 10.12.2006).

III. ANALYSIS USING TRAFFIC SPEED

In this section we adapt statistical methods in order to extract the traffic patterns from the data history. By comparing real-time traffic data to these historical data patterns, this synthesis can be used to identify abnormal traffic conditions.

In the metropolitan area of large cities such as Seoul, the heaviest traffic hours during weekdays are when people travel to and come from work; weekends and holidays, however, the volume of traffic tends to be heavy throughout the day, regardless of business hours. Even on weekdays, the traffic pattern can vary from day to day depending on certain variables. Therefore, we have divided our data into 8 categories: 7 of them correspond to the typical days of the week, with an additional category for holidays.

For the first step of our analysis, we calculated the means of the records of every links for every 10-minute time period in each day's category. The mean value of the data from any day-time combination can give us an idea of the *"normal"* traffic speed of a link at a specific day and time. If the current traffic speed of a link at a certain time of day is much slower than the average, one could suspect that the link might contain an unusual incident.

However, the mere differences between current, actual speed and average, hypothetical speed is not a good enough indication of the abnormality of the traffic flow within a link. For example, let us assume there are two links A and B, with average traffic speeds of 70 km/h and 30 km/h respectively, based on traffic data history. If the measured current speed were 60 km/h for link A and 20 km/h for link B, the differences of current speed and average speed are equally 10 km/h for both A and B. However, it is not fair for one to say that the levels of abnormality of both links are the same.

Therefore, we can adapt the standard deviation as the measure of reasonable statistical dispersion.

$$\frac{C_v - M_v}{\delta_v} \quad (1)$$

C_v is the current velocity of a link, M_v is the average velocity, whereas, δ_v is the standard deviation. In this case, where the current speed is slower than the average speed, the negative numerator result in a negative abnormality which means the case is fairly normal.

Yet, (1) still leaves us with another problem. For example, if the value of a measured current speed of a link subtracted by the average speed is -15 and the standard deviation is 10, the result of (1) will be -1.5. Let us assume another link with -1 as the value of a measured current speed subtracted by the average speed with 0.5 as the standard deviation. In this case, the result of (1) will be -2. Thus, a link with -1 as its value of current speed subtracted by the average speed would end up with a higher abnormality measure.

$$\frac{C_v - M_v}{\delta_v + A\delta} \quad (2)$$

(2) is our solution to the problem presented above. Note that we have added A_δ, the average of the standard deviations of all links, to the denominator. This prevents link records with smaller standard deviations from receiving a higher abnormality. If we apply this notion to our previous example, then assuming the average of all standard deviations is 4, the abnormality measure of the first link is -1.0714, whereas, the other receives -0.2222. Recall that the link with the value furthest from zero is the most abnormal since it deviates the most from the mean—this is regardless of whether the value is positive or negative.

IV. ANALYSIS USING ROAD FEATURES

Other than analyzing the history of traffic speed using statistical approaches, we have also tried to find the hidden relations between various road features (or attributes) and the volume (speed) of traffic. We plan to analyze several distinctive characteristics of the road links, such as the number of lanes, the speed limit, and the kind of vehicles that are restricted from use of that specific road. For example, it seems quite reasonable for one to expect the traffic to be faster on roads with more lanes and higher speed limits. This section presents how we found features that contribute to and affect actual velocity.

In order to indicate the level of influence that each feature has on the traffic speed, we import the concept of ***mutual information***. Mutual information is the reduction in uncertainty about one variable due to the knowledge of the other variables [7].

$$I(X;Y) = \int_Y \int_X p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \quad (3)$$

Where $p(x,y)$ is the joint probability distribution function of X and Y and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively. Larger mutual information means higher dependency, and if X and Y are independent, their mutual information is zero since knowing X does not give any information about Y.

V. Applications

Existing traffic information systems only provide information about a fixed set of major roads on a repetitive basis. We believe that our study will help with developing more dynamic and flexible real-time traffic information systems based on the current, actual status of the road. We introduce new applications implemented on the foundation of our traffic speed and road feature analysis.

A. Prioritized Information Transmission - Traffic Speed Based Method

Our first application utilizes the idea of our traffic speed analysis for the objective of deciding which roads are more important than others (which roads need traffic information more urgently than the others). This system includes a table that contains the average and standard deviation values of traffic speeds of all possible days, times, and link combinations. Using real-time information of the current day, time, and speed of every link on the map, it calculates the abnormality measure following (2). The final output of the program is the rank of links, sorted by their abnormality measures, which essentially tells us the transmission priority of each link.

B. *Prioritized Information Transmission - Road Feature Based Method*

For our second application, we calculated the mutual dependence of various feature values and the speed of traffic, and identify the kinds of links that the system should be transmitting information prior to others.

· Features that count – Out of a total of 17 features, we excluded the features that are used only to distinguish between one another and, therefore, hardly imply any hint about the actual characteristics of the link, such as 'link ID,' 'ROAD_NO,' and so on. Such features are merely different names of the links, which means that they have no impact upon the real and current traffic situation of links.

· Expanded features – More importantly, we have divided all the features into as many briary features as each of feature's possible values. For example, if the feature 'ROAD _USE' had two possible values, 0 (road closed), and 1 (road open), we divided 'ROAD_USE' into two distinct features 'road use 0' and 'road use 1.' A link with a 'ROAD_USE' value of 1 will now have two values of 0 and 1 for 'road use 0' and 'road use 1,' respectively. The expanded feature is a reasonable approach because we are interested in not only which features relate to and impact traffic speed the most, but also the actual feature values that have the most influence on it. Even if we knew, from feature selection, that the feature 'REST_VEH' had the most influence on the traffic speed (which it really did), it is still not enough to say that roads with which value of 'REST_VEH' would actually have more tendency to have heavier traffic. For the features that have continuous values we bind the values into several discrete intervals. Consequently, the original features have expanded to a new set of 99 features.

· Mutual information – Once the original feature set is expanded into its binary version, it is possible to derive the mutual information of every binary feature and the traffic speed. These numbers will indicate how sensitive the traffic is to such feature values. Then, for each link on the map, the mutual information of all the binary features with true values can be added up as an indicator of the link's traffic sensitivity: The higher the mutual information sum is, the more sensitive the traffic is in that link. Thus, a system may consider giving priority to such links. The main idea of this application is to consider links as combinations of feature values. Once we determine which feature values the traffic speed is more sensitive to, it is easy to see that the traffic will have a tendency to change more in the links that are combinations of feature values with higher mutual information.

· Adequacy – In determining the adequacy of our approach, we have compared the prediction accuracy of our system using the top 49, the lower 50, and the total of 99 from the feature ranking according to its mutual information. (See Table II.) The results show that the prediction using the top 49 features was the most accurate, and we believe this proves our link priority approach according to mutual information correct. For the prediction technique, we used Support Vector Machine (SVM) regression from the LIBSVM libraries

TABLE II
ACCURACY OF TRAFFIC SPEED PREDICTION

Day	Time	Total of 99	Top 49	Bottom 50
Mon.	18:30	11.7444	8.2188	12.878
	07:30	12.6230	8.8565	14.0030
Wed.	18:30	11.7765	8.2999	12.9514
	07:30	12.1290	8.4147	14.3780

The root mean square values show that the prediction using the top 49 features (the upper half of the whole ranking) had the best performance.

This result justifies our link priority approach according to mutual information correct.

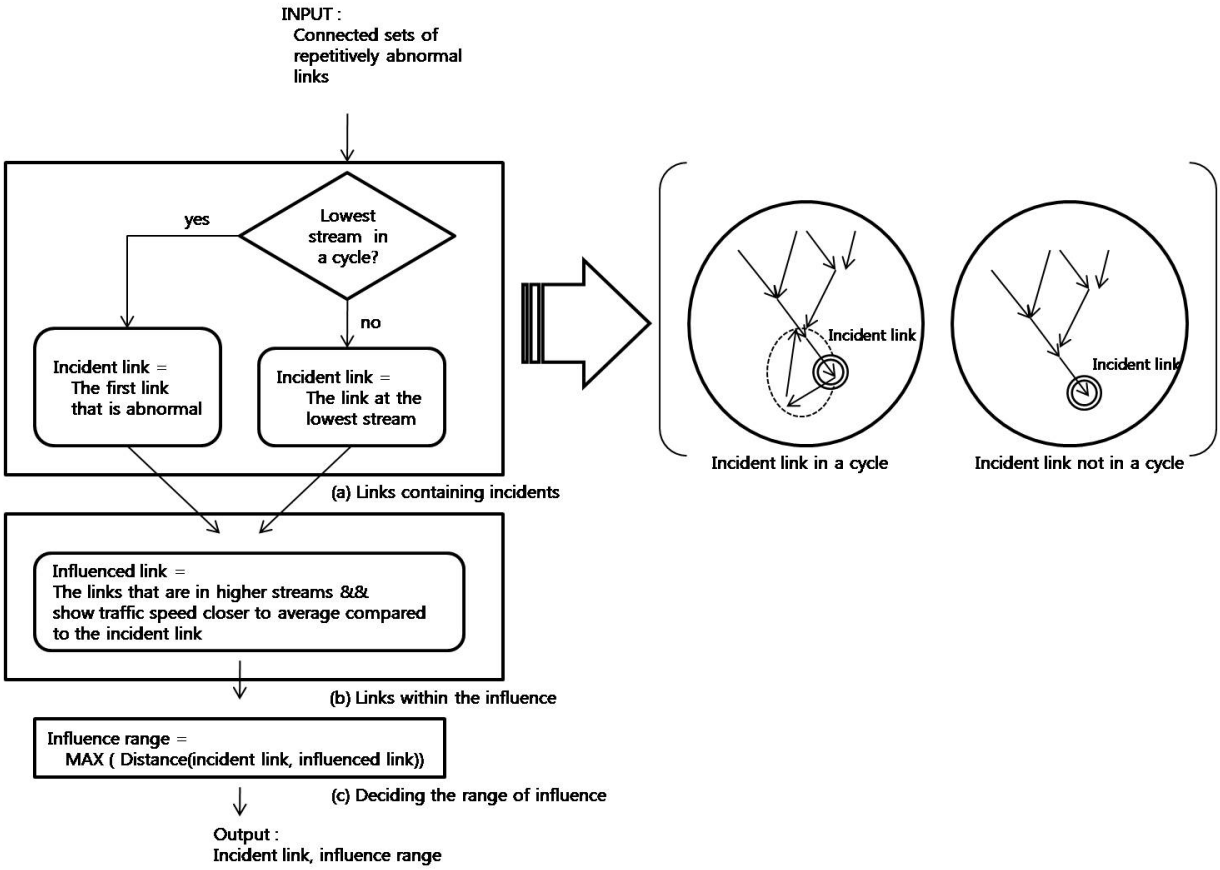

Fig. 1. The process of extracting the instance-influenced traffic data. (a) shows how we found the incident links according to the rule shown on the right image. (b) presents the rules of determining links under the influence of the incident links passed on from (a). In (c), you can see that the influence range of an incident is given as the furthest influenced link from the incident.

[8]-[10].

C. *Predicting the Influence Range of Incidences*

In our study, the road feature based method was adopted as the key to the problem of incident management. As the result, we present a system that is intended to predict the range of traffic congestion caused by an incident. First, we explain the criteria which we relied upon for sorting out the hypothetical instance-influenced records from the whole data history. Then, we show the flow of our system design.

1) *Incident-Influenced Traffic Data:* Since the actual data regarding the effect of incidents was unavailable, a hypothetical dataset was created. To help your understanding we list the procedure we used in the creation.

· Abnormal records – We defined "abnormal records" as link records that show at least 30 km/h lower traffic speed than the average speed of all records at the same time on the same day of the week. The threshold of 30 km/h is a symbolic value of the smallest speed change that people would consider "abnormal." We were able to collect 53,655 abnormal records from the whole data set. The threshold value may vary according to region, kind of road, etc.

· Repeated links – From the set of abnormal data, we sorted out the links that appeared repetitively in continuous 10-minute time slices. This means we have excluded the links that were abnormal for only one time slice and became normal at the very next one. Through this step, we are left with only the links that show abnormal traffic speed for a long enough time period to trigger a case of traffic congestion.

· Connected links – Due to there connection, it is inevitable that the links affect one another, as well, and we sought for the associations among these abnormal repeated links. If the starting node of a link is identical to the end node of another link, both links are connected to each other.

· Links containing incidents – Among the sets of connected links, we select only one link per set as the location of incident. The basic rule we have used was to appoint the link at the lowest stream at the center, the link which contains the incident. However, if the line connection forms a cycle at the lower end, it is impossible to select just one link. In this case, we chose the link which was the first to show congestion (or abnormality). (a) of Fig. 1 shows the flow of this procedure.

· Links within the influence of incidents – After the centers of all connection sets have been decided, we treat the rest as links within the influence of the incident. Other than this main rule, note that we have applied a simple additional regulation to the backtracking procedure: We only choose the links which show reduced current-average speed difference as influenced links. The abnormal links which are connected to

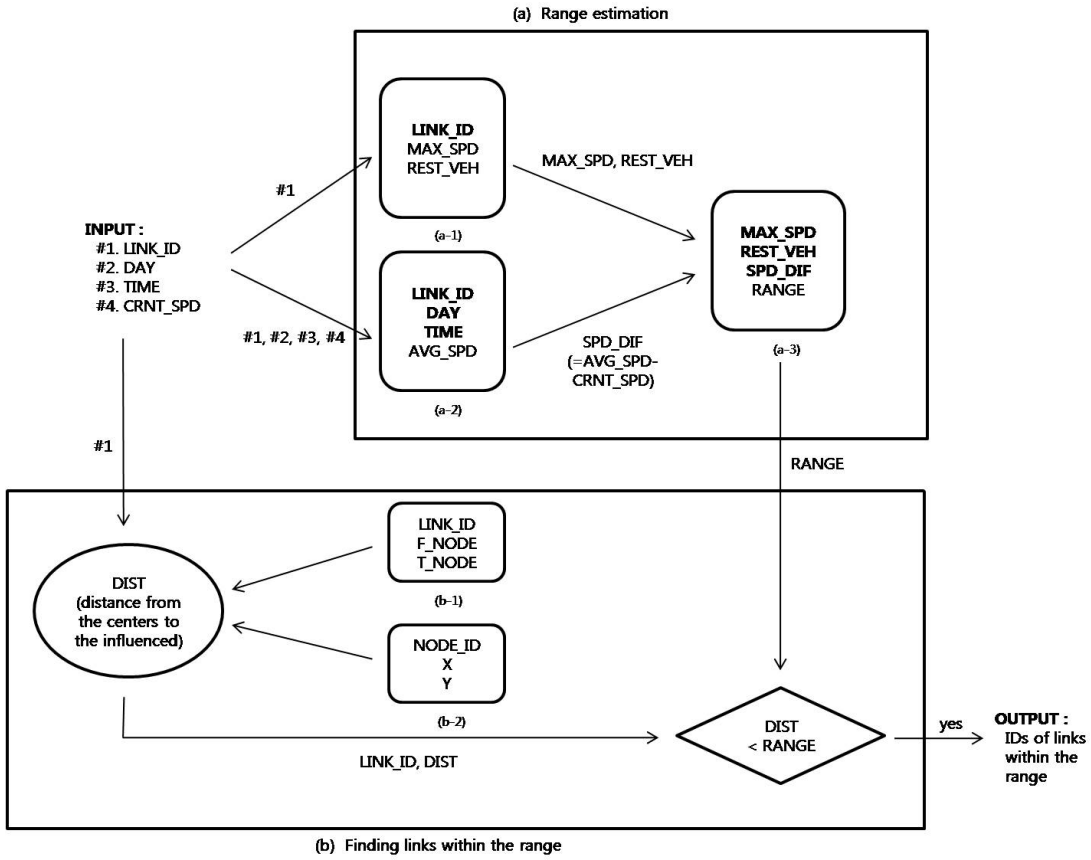

Fig. 2. The flow of predicting the instance influence range is presented. (a) is the program which finds the estimated range of the incident from the range table, where as (b) finds the links within the output range of (a) using the vector information of the link. (a-1), (a-3), (b-1), and (b-2) are tables of the database which contains the necessary feature values of all the links in Seoul area. The contents (names of the features) are written inside each rounded boxes. (a-3) is the range table. According to 3 features of the given link, this table outputs the range value, as the range estimate. In each table, features that are written in bold characters represent the values which are used as the keys. The tables return the queried feature values according to the key values.

the incident links but have larger current-average speed differences are newly registered as incident links. This additional rule arises from our interest in the decrease of the influence based on an area's distance from the incident. (See (b) of Fig. 1.)

· Influence range – We have defined the influence range of an incident as the distance from the center of the incident's location to the furthest link under its influence. (See (c) of Fig. 1.)

2) System Overview: Now we present the overview of our incident influence range prediction system. Remember that the actual incident information and data should replace our hypothetical data set constructed as above.

· Feature selection – As mentioned earlier, the requirement of our system is to derive the influence range of an incident based on the data regarding incident from the past. The road features of the instance links are used as keys in the process. For the sake of handiness, we have selected the top three features that are most relevant to the range of influence, which were speed limit, vehicle restriction, and the standard deviation of speed from mean speed.

· Range table – The most critical part of our system is the range table. The table contains the range averages of every possible combination of features. We have clustered the incident links according to their three feature values, and for each feature combination, we can calculate the mean range of the incident links. (See (a) of Fig. 2.)

· Range of influence – The range table is the summary of the history of incidents. The system uses the feature values of the actual incident link as an index, and returns the average of past records as the influence range estimate. Part (a) of Fig. 2 returns the influence range of a given incident as its output.

Links within the range – Once the system receives the predicted influence range from the table it searches the upper stream of the incident for links that fall in the range, and returns the link IDs to the user. (See (b) of Fig. 2.)

VI. Conclusion

We succeeded in extracting traffic patterns from the statistical analysis on the data history. In addition, we were able to identify a form of relation between the speed of the traffic flow and the features of the road. We were successful in finding how sensitive the traffic is to the features of the road. These two results will form the backbone of our final system as we continue the development process in the future.

We also suggested a solution to the problem of handling incidents. We try to derive the boundaries of the effects caused by the incidents and introduce a new application based on our research.

We believe that our study will provide a solid background, as well as a guide line for the solution systems which regards providing real-time traffic information to the users. Eventually, we aim to construct the prototype for a pattern-based DMB-telematics traffic route guidance system that can lead the way of the user to a more efficient and pleasant path.

In terms of future work, we hope to expand the models with additional training data and evaluation runs. Especially, our instance influence prediction system leaves us with room for further tuning. We intend to experiment with data recorded from actual incidences, and, therefore, eliminate the need for hypothetical incident data. Our approach can also be applied to other measurements of traffic, such as occupancy and volume. Expanding the variety of measurements will give us a deeper understanding of traffic patterns within historical data. It is also necessary that we try diverse selections of the features in the system in order to find the feature combination which shows best performance in practice.

Applications that are capable of traffic speed prediction are also under development. After we achieve enough performance with our systems, we would be able to extent the region nationwide.

REFERENCES

[1] W. Zhu, M. Barth, "Vehicle Trajectory-Based Road Type and Congestion Recognition using Wavelet Analysis," in Proc. IEEE 9th Int. Conf. Intelligent Transportation systems, 2006, pp. 879-884.

[2] A. P. Boedihardjo, C. T. Lu, "AOID: Adaptive On-Line Incident Detection System," in Proc. IEEE 9th Int. Conf. Intelligent Transportation systems, 2006, pp. 858-863.

[3] Y. Chen, Y. Zhang, J. Hu, and D. Yao "Pattern Discovering of Regional Traffic Status with Self-Organizing Maps" in Proc. IEEE 9th Int. Conf. Intelligent Transportation systems, 2006, pp. 647-652.

[4] E. J. Schmitt, H. Jula, "Vehicle Route Guidance Systems: Classification and Comparison" in Proc. IEEE 9th Int. Conf. Intelligent Transportation systems, 2006, pp. 242-247.

[5] A. Guin, "Travel Time Prediction using a Seasonal Autoregressive Integrated Moving Average Time series Model," in Proc. IEEE 9th Int. Conf. Intelligent Transportation systems, 2006, pp. 493-498.

[6] http://road.moct.go.kr/STD_NodeLink_Intro.aspx (Korean).

[7] R. O. Duda, P. E. Hart, D, and D. G. Stork, "Pattern Classification," 2nd ed., Wiley-interscience, 2001, pp. 632-633.

[8] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools (English).

[9] C. H. Wu, J. M. Ho, and D. T. Lee, "Travel-Time Prediction with Support Vector Regression," IEEE Trans. Intelligent Transportation Systems, vol. 5, no. 4, pp. 276-281, DEC. 2004.

[10] S. R. Gunn, "Support vector machine for classification and regression," Tech. Rep., Univ. Southampton, Southampton, U.K., May 1998.