HIGH PERFORMANCE MULTIAGENT

INTERPOLATION SYSTEM by PROBE CAR DATA

Tetsuo Morita, Junji Yano, and Kouji Kagawa
Sumitomo Electric Industries, Ltd.
Information & Communication Laboratories
1-1-3, Shimaya, Konohana-ku, Osaka, 554-0024 Japan
Tel: +81-6-6466-5607, Fax: +81-6-6462-4586
{morita-tetsuo, yano-junji, kagawa-kouji}@sei.co.jp

ABSTRACT

We propose a multiagent interpolation system for traffic conditions that includes estimation and learning agents. These agents are allocated to all the road links. The Normalized Velocity (NV) is used in this system. Estimation agents renew the NV for each road link, and learning agents renew the weight values for estimation. The weight values can be calculated by multivariate analysis. Estimation and learning agents alternately calculate the results to improve the interpolation accuracy. The coefficient of determination and mean square error are used to evaluate the interpolation accuracy. The obtained probe car data is stored in a learning database. In this study, we change a hard disk drive-based (HDD-based) learning database into a memory-based one to improve the performance. The processing time is reduced to 1/30, when the memory-based learning database is used. The average standard deviation of the estimated velocity error is 7.34 km/h in the evaluation area.

INTRODUCTION

The role of traffic information services in reducing the consumption of fossil fuels and carbon dioxide is important. Traffic information is classified into two types: temporal information and spatial information. Temporal information denotes forecast technologies, and there is no effective way to forecast future traffic congestion. Spatial information corresponds to traffic congestion mapping. The multiagent interpolation system we propose can estimate traffic conditions based on small amounts of information.

The Vehicle Information and Communication System (VICS) is well known as a traffic information service. VICS gathers the traffic information from roadside sensors and provides it to drivers. VICS is very useful in providing traffic information, but a huge capital investment in roadside sensors is essential. The probe car system is an effective method of reducing this capital investment. Probe cars measure the travel time along road links using
Global Positioning System (GPS) sensors and other methods. With the probe car system, it is not necessary to install sensors at the roadside. However since probe cars are very few in number, it is difficult to estimate traffic congestion only from Probe Car Data (PCD).

One commonly-used method to interpolate traffic conditions is by statistical analysis. This method uses statistical data for the time-sliced average of past PCD from road links. At present, the number of probe cars providing data that covers the same time and conditions is very few, and the sampling errors are very large.

The pheromone model [1] is used to make up the deficit in PCD, with deposit, propagation, and evaporation as the pheromone parameters. While the pheromone model is normally used as a forecast technology, it can be used for interpolation. The pheromone intensity depends on the velocity of the traffic, and changes through a mechanism of propagation and evaporation. Thus traffic congestion can be estimated from the pheromone intensity. But the pheromone parameters are determined by human experience and are difficult to determine objectively.

The Feature Space Projection (FSP) method [2][3] is proposed as a method to interpolate the traffic conditions, with the feature being obtained by Principal Component Analysis (PCA) with missing data. The PCA method without missing data is commonly used for multivariate analysis. But in this case, the probability of getting simultaneous PCD from two road links is very small, and so a method of using PCA with missing data is essential for this calculation.

A multiagent interpolation system [4][5] for traffic conditions, that includes estimation and learning agents is proposed. These agents are allocated to all the road links. Estimation agents renew the velocity for each road link, and learning agents renew the weight values for estimation. The weight values can be calculated by multivariate analysis. Estimation and learning agents alternately calculate the results to improve the interpolation accuracy.

In this paper, we improve the performance of the multiagent interpolation system. Learning agents calculate the weight value by the velocity of the reference road links, which are stored in the learning database. As a Hard Disk Drive (HDD) is used for the learning database, the HDD access is the bottleneck in this system. The size of the learning database is too large to store it in the memory of a PC. However the velocities of the road links can be stored in the memory of the PC when compressed, which greatly improves the performance of the system.

**MULTIAGENT INTERPOLATION SYSTEM FOR TRAFFIC CONDITIONS**

Figure 1 shows the multiagent interpolation system for traffic conditions. This system consists of both estimation and learning agents that are allocated to all the road links. Estimation agents renew the velocity for each road link, and learning agents renew the weight values for
estimation. Estimation and learning agents alternate in calculating the results to improve the interpolation accuracy. The estimated velocities and the weight values are stored in the velocity/weight database. The velocities for learning agents are selected and copied in the learning database. In this study, we change an HDD-based learning database into a memory-based one to improve the performance of the system. The upper limit on the number of PCD values stored in the learning database is variable. The normalized velocity (NV) used in this system is given by $y = 1 - x/100$, where $x$ denotes the velocity (km/h), and $y$ denotes the NVs. When the $x$ is more than 100 (km/h), 0 is assigned to $y$.

![Schematic Diagram of Interpolation system for Traffic Conditions](image)

Estimation agents calculate the NV for the road link at time $t$ which is estimated using the NVs for the reference road links and the weight values at time $t-1$, and the reference road links are adjacent to the road link being estimated. The initial NV for each road link is 0, and the initial weights are $w_0^{(i)} = 0$, $w_1^{(i)} = \cdots = w_n^{(i)} = 1/n$. Subscript $n$ denotes the number of reference road links, and each road link has a different value of $n$. While the notation $n^{(i)}$ is strictly correct, the superscript $i$ for the $n$ is omitted. Without initial values, the multiagent interpolation system cannot break the deadlock.

$\mathbf{V}^{(i)}$ denotes the NV vector for the reference road links associated with road link $i$, and $w^{-1} \mathbf{w}^{(i)}$ is the weight vector of the $i$-th road link at time $t-1$. $\mathbf{V}^{(i)}$ consists of $n$ NVs for the reference road links and a constant value 1. Equation (1) shows the definition of the estimated NV $\tilde{E}^{(i)}$ for road link $i$ at time $t$. In other words, the estimated NV is the inner product of the NV vector and the weight vector. Occasionally, the weight value $w_0^{(i)}$ is referred to as the threshold.

$$
\tilde{E}^{(i)} = \mathbf{V}^{(i)} \cdot w^{-1} \mathbf{w}^{(i)} = \begin{pmatrix} v_0^{(i)} & \cdots & v_n^{(i)} \end{pmatrix} \begin{pmatrix} w_0^{(i)} & \cdots & w_n^{(i)} \end{pmatrix} = \mathbf{v}^{(i)} (w^{(i)}) = \begin{pmatrix} \cdots \end{pmatrix} $$

(1)
Figure 2 shows an example of the road link connections. The travel time for each road link is converted to an NV, and the NVs at time $t-1$ are $t-1\nu^{(1)}, \ldots, t-1\nu^{(7)}$. $t\nu^{(1)}$ is the NV for road link 1 at time $t$. When the road link being estimated is 2 and the reference road links are 1, 3, and 6, $t\hat{E}^{(2)}$ can be calculated from Equation (2).

$$t\hat{E}^{(2)} = \begin{pmatrix} t\nu^{(1)} & t\nu^{(3)} & t\nu^{(6)} & t^{-1}w_0 & t^{-1}w_1 & t^{-1}w_2 & t^{-1}w_3 \end{pmatrix}^T$$ \tag{2}

Learning agents are allocated to all the road links, and they calculate the weight vector $w$ for learning road link $i$ referring to the PCD NVs for road link $i$ and the NVs for the reference road links. The superscripts for time $t$ and road link number $i$ are omitted for the learning agents. Equation (3) is $m$ simultaneous equations in $n + 1$ unknowns. $P$ denotes the PCD NV vector for road link $i$ with $m$ NVs, and $V_{mx(n+1)}$ denotes the matrix consisting of $m$ NV vectors for the $n$ reference road links and $m$ constant values of 1. Subscript $m$ denotes the number of PCD values, and each road link has a different value of $m$. While the notation $m^{(i)}$ is strictly correct, the superscript $i$ for $m$ is also omitted.

$$P = V_{mx(n+1)} \cdot w$$ \tag{3}

$$P = (P_1 \cdots P_m)^T$$ \tag{4}

$$V_{mx(n+1)} = \begin{pmatrix} 1 & V_{11} & \cdots & V_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & V_{m1} & \cdots & V_{mn} \end{pmatrix}$$ \tag{5}

When $m$ is less than $n + 1$, the solutions to the simultaneous equations in Equation (3) are not fixed. When the rank of $V_{mx(n+1)}$ is $n + 1$, Equation (3) can be solved. If the number of independent equations is greater than $n + 1$, Equation (3) cannot be solved. In this case, the
least mean squares method can be used to minimize the Mean Square Error (MSE). The dependent variables are the PCD NVs, and the independent variables are the NVs of the reference road links. Equation (6) can be derived from Equation (3) [4]. Equation (6) can be solved by the Gaussian elimination method, and the regression coefficients \( w_1^{(i)}, \ldots, w_n^{(i)} \) can be calculated.

\[
\begin{pmatrix}
  s_{11} & s_{12} & \cdots & s_{1n} \\
  s_{21} & s_{22} & \cdots & s_{2n} \\
  \vdots & \vdots & \ddots & \vdots \\
  s_{n1} & s_{n2} & \cdots & s_{nn}
\end{pmatrix}
\begin{pmatrix}
  w_1 \\
  w_2 \\
  \vdots \\
  w_n
\end{pmatrix} =
\begin{pmatrix}
  p_1 \\
  p_2 \\
  \vdots \\
  p_n
\end{pmatrix}
\]  

(6)

\[
s_{qr} = \frac{\sum_{j=1}^{m} (V_{jq} - \bar{V}_q)(V_{jr} - \bar{V}_r)}{m-1}
\]

(7)

\[
p_q = \frac{\sum_{j=1}^{m} (P_j - \bar{P})(V_{jq} - \bar{V}_q)}{m-1}
\]

(8)

\[
\bar{V}_q = \frac{\sum_{j=1}^{m} V_{jq}}{m}, \quad \bar{P} = \frac{\sum_{j=1}^{m} P_j}{m}
\]

(9)

**PERFORMANCE OF PROCESSING**

PCD for Nagoya taxis is used in this evaluation. The evaluation area for this system is the Nagoya district including Nagoya Station, an area of approximately 10 kilometers by 10 kilometers on a longitude from 136°52’30” to 137°00’00” and a latitude from 35°10’00” to 35°15’00”. The taxi company that took part in this experiment has approximately 1200 taxis, and the total number of road links is 1128. The PCD can be obtained every 15 minutes. The PCD from 01/11/2007 to 31/10/2008 was used for this evaluation.

First, the number of PCD values from the 1128 road links was checked (see Table 1). The subtotal for road links with single-figure PCD values is 87. We should note that the number of PCD values \( m \) must be not less than the number of unknowns \( n+1 \) for learning; therefore it is difficult to calculate the weight value of a road link from single-figure PCD values. The subtotal for road links with four- and five-figure PCD values is 348. In this range, the road links have collected a sufficient number of PCD values. The number of PCD values obtained per day was 96, and the evaluation period was 366 days, giving a maximum number of PCD values of 35136.
In the four- and five-figure ranges, the NV for the road links could be estimated by traditional methods without using this interpolation system. At the end of the evaluation period (31/10/2008), the number of road links with PCD was 1074 at most. The rest of the 54 road links had no PCD. A suitable range of PCD for this interpolation system is considered to be three figures.

Table 1. Frequency Distribution of Number of PCD Values

<table>
<thead>
<tr>
<th>Number of PCD</th>
<th>Count</th>
<th>Number of PCD</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19</td>
<td>10 To 19</td>
<td>59</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>20 To 29</td>
<td>32</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>30 To 39</td>
<td>34</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>40 To 49</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>50 To 59</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>60 To 69</td>
<td>20</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>70 To 79</td>
<td>23</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>80 To 89</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>90 To 99</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Subtotal</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Subtotal</td>
<td>242</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of PCD</th>
<th>Count</th>
<th>Number of PCD</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 To 199</td>
<td>107</td>
<td>1000 To 4999</td>
<td>237</td>
</tr>
<tr>
<td>200 To 299</td>
<td>52</td>
<td>5000 To 8999</td>
<td>58</td>
</tr>
<tr>
<td>300 To 399</td>
<td>53</td>
<td>9000 To 12999</td>
<td>22</td>
</tr>
<tr>
<td>400 To 499</td>
<td>38</td>
<td>13000 To 16999</td>
<td>12</td>
</tr>
<tr>
<td>500 To 599</td>
<td>32</td>
<td>17000 To 20999</td>
<td>5</td>
</tr>
<tr>
<td>600 To 699</td>
<td>34</td>
<td>21000 To 24999</td>
<td>6</td>
</tr>
<tr>
<td>700 To 799</td>
<td>30</td>
<td>25000 To 28999</td>
<td>7</td>
</tr>
<tr>
<td>800 To 899</td>
<td>23</td>
<td>29000 To 32999</td>
<td>1</td>
</tr>
<tr>
<td>900 To 999</td>
<td>28</td>
<td>33000 To 36999</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Subtotal</td>
<td>397</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Subtotal</td>
<td>348</td>
<td></td>
</tr>
</tbody>
</table>

| Sum Total         | 1074  | No PCD               | 54    |

The PCD NV vector \( P \) and the NV matrix \( V_{\text{max}(n+1)} \) are stored in the learning database. The size of the learning database increases in proportion to the number of PCD values \( m \). The size of the learning database was approximately 500MB at the end of the evaluation (31/10/2008). When new PCD is obtained (1 added to \( m \)), the values of \( P \) and \( V_q \) in Equation (9) are
changed. Therefore we have to recalculate the values of $s_{qr}$ and $p_q$ in Equations (7) and (8).

The number of coefficients $s_{qr}$ and $p_q$ in Equation (6) is $n^2 + n$. As the matrix in Equation (6) is symmetrical, the number of coefficients that should be calculated is $n(n+1)/2 + n$. When the number $m$ increases, the number of coefficients no longer depends on $m$. Equations (10) and (11) are the transforms of Equations (7) and (8). The coefficient of Equation (6) can be calculated by the sum of $V_{jq}$, $V_{jr}$ and the sum of the product of $V_{jq}$, $V_{jr}$ and $P_j$, $V_{jq}$.

The size of the learning database drastically decreases when Equations (10) and (11) are used.

$$s_{qr} = \frac{m \sum_{j=1}^{m} (V_{jq} \cdot V_{jr}) - \sum_{j=1}^{m} V_{jq} \sum_{j=1}^{m} V_{jr}}{m(m-1)}$$

(10)

$$p_q = \frac{m \sum_{j=1}^{m} (P_j \cdot V_{jq}) - \sum_{j=1}^{m} P_j \sum_{j=1}^{m} V_{jq}}{m(m-1)}$$

(11)

Figure 3 shows the processing time of the interpolation system from 01/11/2007 to 31/10/2008. The HDD-based learning database takes approximately two weeks for calculation, but the memory-based one takes only 12 hours; the processing time was reduced to 1/30, when the memory-based learning database was used. For calculation, an Intel Core2 Duo CPU (clock speed 1.2 GHz, 1GByte RAM) was used.

![Figure 3. Processing Time of Interpolation System](image-url)
EV ALUATION by CD and MSE

In this simulation, the Coefficient of Determination (CD) and MSE are used to monitor the progress of learning [4]. Figure 4 shows the fluctuations in the CD for several road links in the evaluation area. The x-axis denotes the elapsed time in days, and the y-axis denotes the CD. The PCD used for this evaluation was from 01/11/2007 to 31/10/2008. The same PCD was used twice. Firstly, the PCD from 01/11/2007 to 31/10/2008 was used, and the same PCD from 01/11/2007 to 31/03/2008 was used thereafter. The road links for which the number of PCD values is between 2500 and 3000 are selected. The upper limit on the number of PCD values stored in the learning database is 32000 for each road link. Therefore no PCD values are removed in this evaluation.

The fifth digit in the road link number (e.g. 30567) denotes the following:

1: Inter-city Expressways; 2: Inner-city Expressways; 3: Local roads

The rest of the digits denote the number assigned to the road link. The road link number in Figure 4 denotes a local road.

![Figure 4. Fluctuation in the Coefficient of Determination](image)

When the number of PCD values is less than the number of unknowns \((m < n + 1)\), 0 is assigned to the CD. When Equation (3) can be solved, the CD becomes 1. As the number of PCD values increases, the CD decreases abruptly, and then increases gradually. The minimum value of the CD for road link 30245 is approximately 0.2 after learning, and its value gradually increases to 0.82. The CD for road link 31132 abruptly decreases when the number of PCD values is approximately 1500, and increases gradually thereafter. The reason why the CD changes abruptly is thought to be that the weight values of the reference road links change. The reason why the CD increases gradually is thought to be that the PCD is concentrated in a hyperplane in the multivariate analysis.
The weight values can be calculated by multivariate analysis, which minimizes the MSE, which is the essential parameter in the evaluation. Figure 5 shows the fluctuations in the MSE. The x-axis denotes the number of PCD values, and the y-axis denotes the MSE. When no PCD exists for the road link, 1 is assigned to the MSE. When Equation (3) can be solved, the MSE becomes 0. As the number of PCD values increases, the MSE increases abruptly, and then decreases gradually. The trend in the MSE is opposite to the CD.

Figure 5. Fluctuation in the Mean Square Error

Figure 6 shows the fluctuations in the CD for road links 30587. The x-axis denotes the date, and the y-axis denotes the CD. The evaluation period was from 01/11/2007 to 31/10/2008 in the first run, and the same PCD values from 01/11/2007 to 31/03/2008 were used in the second run. The upper limits on the number of PCD values stored in the learning database are 1000 and 32000 respectively.

Figure 6 indicates that the CD increases when the same PCD values are used twice. When learning is in progress, the weight values of the reference road links are renewed. Even if the road link being learned obtains the same PCD values, estimation agents calculate the NVs for the reference road links with different weight values. Learning agents use different NVs in the multivariate analysis. Since the same PCD values are used twice, the number of PCD values increases substantially. When the number of PCD values is m and the same data is used twice, the actual number of PCD values becomes 2m. The estimation errors during the first run are very large, and therefore iterative use of the same PCD is preferable.

The CD abruptly decreased during New Year’s Day, in mid February (three consecutive national holidays), at the end of April (successive national holidays), in mid August (Bon Festival), and in mid September (national holidays). During these periods, unexpected traffic congestion occurs in this area. When the upper limit on the number of PCD values stored in the learning database is 1000 and the number of PCD values exceeds 1000, the oldest PCD
values are deleted from the database. As the PCD associated with the traffic congestion does not exist in the learning database, the CD decreases abruptly. When the number of PCD values stored in the learning database is 32000, it includes the PCD associated with traffic congestion. Therefore the fluctuations in the CD become moderate. The number of PCD values obtained from the road link is 40169 at the end of the evaluation. The number of removed PCD values is approximately 8000.

![Figure 6. Fluctuations in the Coefficient of Determination](image)

Figure 6 shows fluctuations in the MSE when the same PCD values are used. The x-axis denotes the date and the y-axis denotes the MSE. Figure 7 indicates that the trend in the MSE is opposite to the CD. The MSE for road link 30587 is 0.0068 at the end of the evaluation.

![Figure 7. Fluctuations in the Mean Square Error](image)
Figure 8 shows the frequency distribution of the estimated velocity error of all the road links in the evaluation area. The x-axis denotes the standard deviation of the estimated velocity error, and the y-axis the count. The total number of road links is 1128, and the number of road links without PCD is 65 at 31/03/2008 in the first run. The number of road links without PCD is 54 at 31/03/2008 in the second run. All road links without the PCD are removed from Figure 8.

![Figure 8. Frequency Distribution of Estimated Velocity Error](image)

The dashed line denotes the frequency distribution for the first run (01/11/2007 - 31/03/2008), and the solid line denotes the frequency distribution for the first (01/11/2007 - 31/10/2008) and second runs (01/11/2007 - 31/03/2008). At a position on the x-axis of 2, the standard deviation $x$ of the estimated velocity error is $1 < x \leq 2$. The count is 13 for the first run when the position on the x-axis is 0. This means that the standard deviation of the estimated velocity error of 13 road links is 0, because Equation (3) can be solved.

The average of the standard deviation of the estimated velocity error is 8.37 (km/h) at 31/03/2008 in the first run. The average of the standard deviation of the estimated velocity error becomes 7.34 (km/h) at 31/03/2008 in the second run. The standard deviation of the estimated velocity error decreases when the number of PCD values increases.

**CONCLUSIONS**

We propose a multiagent interpolation system for traffic conditions that includes estimation and learning agents. These agents are allocated to all the road links. Estimation agents renew the NV for each road link, and learning agents renew the weight values for estimation. The weight values can be calculated by multivariate analysis. The interpolation accuracy can be improved when the number of PCD values increases.
We change an HDD-based learning database into a memory-based one to improve the performance. When the memory-based learning database is used, the processing time becomes 1/30. When the number of PCD values stored in the learning database is 1000, there are large fluctuations in the CD and MSE, but when the number of PCD values stored in the learning database is 32000, the fluctuations in the CD and MSE becomes moderate. The average standard deviation of the estimated velocity error becomes 7.34 km/h in the evaluation area.

REFERENCES


