Department of Computer Science, Faculty of Engineering, The University of Hong Kong

FYP17009

Hong Kong Traffic Data Visualization and Prediction

Interim Report as of 21st January, 2018

Supervised by Dr. Choi Loretta

Prepared by WONG CHUN MAN, 3035179393

Abstract

Although online map applications are popular and user-friendly nowadays, most of them only provide static content. Some powerful map displays the traffic speed but the inaccuracy leads to a concern. In addition, there are few map applications that have forecasting function in it.

In this project, how the dynamic map is implemented and the forecast for traffic speed with high accuracy are stressed out. The final deliverable will be a web application which displays the instant traffic speed map and the forecast results of the roads.

The live traffic display is completed and its implementation is explained in this report. The study of forecasting is in progress and streamlining the process is the top priority in the forecasting part. The display output will be refined since and the color segments representing the traffic speed cannot be displayed properly. A backup program should be considered for filling in the missing data. In the forecasting section, some models fit to the data but the identification process is tedious. It is substantial to identify the model in short time and therefore machine learning is proposed.

Acknowledgement

I would like to express my deepest gratitude to my supervisor Dr. Y. K. Choi for giving support for the final year project. I would also like to thank Mr. Ken Ho for his helpful and valuable suggestions on the reports and presentations. Lastly, I want to thank the CS department and Mr. Jonathan Wong for the supports.

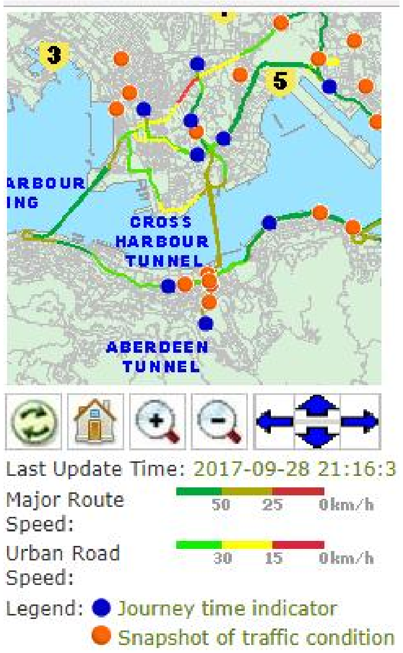
Table of Content

List of Figures

List of Tables

1. Introduction

The Traffic Speed Map providing the estimated traffic speed of major routes to cross harbor tunnels and from the New Territories to Kowloon was launched in 2010 [1]. Users can know the special traffic arrangement, snapshots of roads and the traffic speed map updated for every 5 minutes. Although the sources are reliable, the poor interface of the application leads to low popularity. The fixed small display size (450 pixels x 250 pixels) does not suit the modern screen display. To adjust the map view, users have to use the control panel but not some usual gestures. The magnification and moving scale are also fixed and users cannot view freely (Figure 1). The outdated and non-user friendly application faded out.


3
5
ARBOUR
ING
CROSS
HARBOUR
TUNNEL
ABERDEEN
TUNNEL

Last Update Time: 2017-09-28 21:16:3

Major Route Speed: 50 25 0km/h
Urban Road Speed: 30 15 0km/h
Legend: ● Journey time indicator
● Snapshot of traffic condition

Figure 1 - The traffic speed map available in Transport Department, HKSAR

There are several online map applications available which have better interface then the Traffic Speed Map. Among those applications, Google Map functionality is superior to the other. Google Map displays the elements dynamically with user interaction, for instance, the current traffic situation and routing service. Despite the outstanding performance, there are some deficiencies. Google Traffic works by analyzing the GPS-determined locations transmitted by a large number of users. This method becomes a privacy issue. Secondly, although Google Map provides the forecast for all routes, the speed is undetermined because it is represented in spectrum and the sources are inaccurate. The display often shows the congestion level is high in some non-major routes.

In this project, a web application similar to Traffic Speed Map will be developed. Users can view the current traffic speed level shown in different colors. In addition, some online map applications may estimate the time of certain route based on the current traffic levels. However, it is insignificant to use the current data to estimate the time spent on driving in future. Hence, the application will estimate the traffic speed in future for accurate time estimation.

The system architecture and designs are then introduced with technical justifications. The major components of front-end and back-end development in the architecture are explained in details. After integrating the components, the live traffic displayed is completed and the performance is evaluated. Some alternatives are proposed for some unsatisfactory results. Some results on forecasting are also revealed.

2. Objectives

The objective of the project is to develop a web application which:

- in general,
  - Displays the real-time traffic condition of major routes
  - Updates the display regularly
- for specific roadway,
  - Displays the change of the traffic speed within a time interval
  - Provides information and current condition snapshot
  - Analyses the historical data and predicts the speed.

The construction of the web-application will be carried out in two stages, corresponding to two semesters respectively. By the end of the first stage, the traffic speed of all roads should be visualized in a map correctly and updated regularly. For the second stage, the detailed analysis will be completed.

3. Scope

It is impossible to complete the whole traffic speed map in Hong Kong because of insufficient dataset. Using dataset from Google requires premium plan and acquiring the plan exceeds the project budget. Therefore, we will only focus on the data available in government site.

Web browser is chosen for the project platform as the application requires data request from the server and the database. In addition, most computers, Android and iOS have browser such that no additional skills on developing Android app or iOS are required, which simplify the process.

4. Methodology

This section aims to introduce how the application is implemented. The design of the application is introduced first. Then, the front-end, back-end development and the database structure are stated.

4.1. System architecture

The current preference is to apply the multi-tier server-client architecture which consists of a relational database using SQL, application and web server with PHP support (Figure 2)

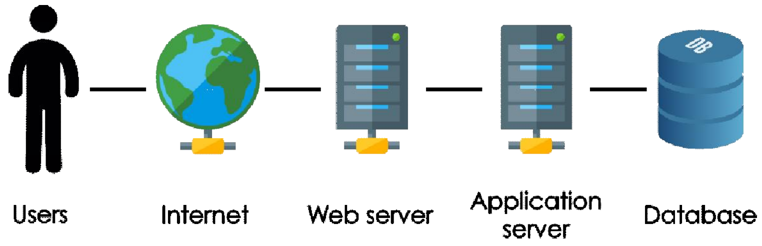


Figure 2 - Multi-tier architecture of the web application

4.1.1. Web server

Web server is responsible for front-end development. This server can be accessed via browsers and links between users and the application server. The server sends requests to the application server for retrieving information and visualizes the data received.

4.1.2. Application server

The server aims to handle requests from the web server and manage data in the database. When receiving request from web server, the server executes SQL statements to retrieve data from database and reformats the results into map-compatible type. Moreover, the server requests data from the government site automatically and periodically to capture the latest data. Since the government does not allow cross-domain request [2], the server is set to be accessible from the

Internet. In addition, the server is implemented in a virtual machine server because auto data request needs to be run constantly.

4.1.3. Hardware & software

The following items used to build up the server are necessary for the project:

- Web server: i.cs.hku.hk/~fyp17009, installed with PHP 5.5.9 and Apache HTTP server 2.4.7
- Application server: fyp17009s1.cs.hku.hk, Window server 2012
- Database: sophia.cs.hku.hk, installed with MySQL database server 5.1.35

Since all the items are available in CS department and can be accessed all the time, the above hardware is adapted. All the hardware was applied and set up already. The servers can be accessed through Internet to ensure the public can access the web application.

4.2. Visualization

This section aims to introduce the stages completed in the visualization part. The data structure of the record is introduced first and the method coordinates conversion is stated. The steps of implementation are mentioned with technical justifications.

4.2.1. Data structure from the government

The data from government site are in XML format. Each route is a node with 6 attributes. Table 1 shows the attributes of each route and its meaning.

<table>
<tr><td>Attribute</td><td>Meaning</td></tr>
<tr><td>Link ID</td><td>The starting and ending nodes, concatenated with “-”</td></tr>
<tr><td>Region</td><td>Region that the road located
<table>
<tr><td>Short form</td><td>Full name</td></tr>
<tr><td>K</td><td>Kowloon</td></tr>
<tr><td>ST</td><td>Sha Tin</td></tr>
<tr><td>TM</td><td>Tuen Mun</td></tr>
<tr><td>HK</td><td>Hong Kong Island</td></tr>
</table></td></tr>
<tr><td>Road type</td><td>Type of the road, either urban road or major route</td></tr>
<tr><td>Road saturation level</td><td>The congestion level, indicated by:
good/ average/ bad</td></tr>
<tr><td>Traffic speed</td><td>The speed per 5 minutes</td></tr>
<tr><td>Capture Date</td><td>The date and time that the speed is captured at</td></tr>
</table>

Table 1 - Attributes in record

For the Link ID, the geographical coordinate information of the starting and ending nodes is provided [3]. The coordinates of starting and ending nodes are represented in HK1980 Grid System. Although formulae are provided for conversion [4], the government has provided the transformation tool with high accuracy. Therefore, a batch conversion can be proceeded.

4.2.2. Front-end development

The presentation layer is implemented in the web server for displaying the data. Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), JavaScript are used for web interface development. HTML is used for creating the web pages and implementing the static information of the website. The website layout is improved by CSS and the dynamic displays are updated regularly by calling JavaScript function. For every minute, the JavaScript function sends as AJAX request to application server for update. Once the request is success and receives the data, the data is loaded into JavaScript function for visualization.

To facilitate the development, online map application tools are used for map display. Google MAP API is chosen because of its documentation and popularity. The documentation provides explicit user guide manual with examples and enhances the development process. Compared with other map applications, users are more familiar with Google Map [5].

4.2.3. Back-end development

The data access layer is implemented in the application server. To handle requests, manage database access and request data from the government site, PHP, a server-side scripting language for web development, is used. Two PHP files are implemented. One is for web development and responding requests from web server. The other one is an auto request data program running continuously to capture data from government.

4.2.3.1. Auto data request program

Because the data in the government website updates for every 1 – 3 minutes, the program sends a HTTP GET request to the government site for every minute. Upon successful request, some verification is conducted to ensure no duplicate data is inserted. Figure 3 shows the overview process of inserting data into database. If data request succeed, 3 random data are selected and checked if each capture time is same with the latest capture time of that road retrieved from database. If the results are the same, it implies the data in the government site is not update, then server waits another minute. If all results are different, then the server inserts the latest record and updates the time. Otherwise, 3 more roads are selected and repeat the above process.

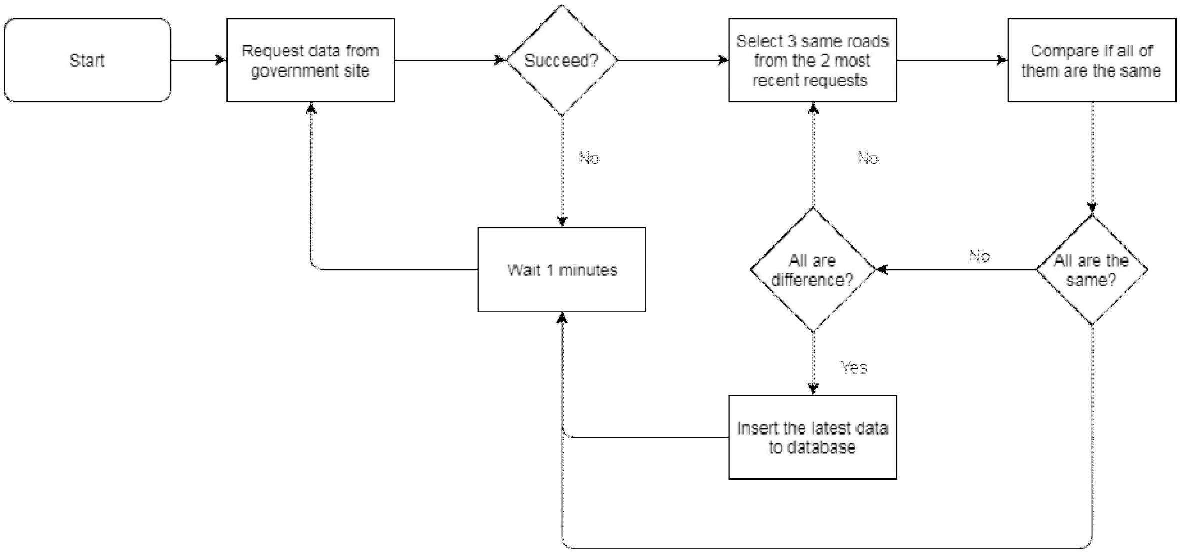

Start
Request data from government site
Succeed?
Select 3 same roads from the 2 most recent requests
Compare if all of them are the same
No
No
Wait 1 minutes
All are difference?
No
All are the same?
Yes
Insert the latest data to database

Figure 3 - Flowchart of the automated request and the verification process

4.2.4. Database Design

Data received from the government site can be summarized as two categories, static and variable. For instance, Road ID, geographical location, and type of route are static while speed and capture time are variable. Therefore, all static information can be stored in a table while the variables of each route should be stored in one table.

To facilitate the forecast process, the structure of the tables should reduce the searching time and minimize the number of tables searched. To uniquely identify a record, the keys are: Date of the data recorded, Time that the data retrieved, and the Road ID of specific road. Some keys are selected as the table name and the others are used for the primary keys of that table. Since the Time varies widely. It should not be used in the table name. Using Road ID as a table name and Date and Time as primary key can reduce the number of tables created after certain period.

4.3. Forecasting

In this session, the methodology of the forecasting is presented. We first pre-process the data and identify suitable models for fitting the series. After identifying the model and its corresponding parameters, we will use the model to forecast the next day dataset.

4.3.1. Data preprocessing

Since the government does not update the records regularly, the data forms an irregular time series increasing the complexity of forecasting process. Linear interpolation is adopted because of its simplicity and acceptable forecast results using this method. Polynomial interpolation or nearest-neighbor interpolation are not considered since the points interpolated are not better and the error of the forecast result is significant. Therefore, linear interpolation is used.

4.3.2. Model identification

Some time series models are selected for fitting the data. The selection criterion is whether the model can estimate the model with high accuracy without overfitting. If multiple models are selected, the model with the least number of parameters is selected.

4.3.2.1. Model selection

Autoregressive integrated moving average model (ARIMA) model is selected for analyzing the time series. Compared with the linear trending and exponential smoothing, the model fits the series the best and the forecast results is acceptable.

4.3.2.2. Parameters estimation

ARIMA models are denoted ARIMA(p,d,q) in general, where p and q are the number of parameters in autoregressive and moving average part respectively. d is the number of differencing required to change the non-stationary time series to stationary. If the time series have season trending, it can be modeled as ARIMA(p,d,q)X$(P,D,Q)_s$, where P,D,Q are parameters of seasonal trending with periods.

4.1.1.1.1. Differencing

Since the mean and the variance of the non-stationary time series vary over time and ARIMA model fit the stationary time series only, we difference the time series data [6]. Differencing stabilizes the mean and deseasonalizes the trend. In addition, logarithm is used for stabilized the variance. If we assume the series data consists of a variable that is constant for every two consecutive time points, than we may predict the variable by differencing.

4.1.1.1.2. Identifying the orders of AR and MA

Autocorrelation function (ACF) and partial correlation function (PACF) are used for identifying the orders of MA and AR respectively. Autocorrelation is the correlation between a value and its previous value while partial correlation between two variables is the amount of correlation between them which is not explained by their common variables. The formulae for the ACF and PACF are:

$$ACF: r_k = \frac{\sum_{t=k+1}^{n}(Z_t - \bar{z})(Z_{t-k} - \bar{z})}{\sum_{t=1}^{n}(Z_t - \bar{z})^2}$$

$$PACF: \hat{\varphi}_{kk} = \frac{r_k - \sum_{j=1}^{k-1} \hat{\varphi}_{k-1,jk} r_{j-j}}{1 - \sum_{j=1}^{k} \hat{\varphi}_{k-1,jk} r_j}$$

Where k is interval apart between 2 data.

If there is a sharp peak that against the critical region in ACF or PACF play at lag k, it can be concluded that the MA(k) or AR(k) can fit the series respectively.

4.1.2. Model refinement

After fitting the model, we compute the ACF and PACF of the residual and observe if there are significant peak in the plots. If positive, we increment the complexity of the suggested model and repeat the above process. Else, the model is adequate enough.

5. Results and Evaluations

5.1. Visualization

This session shows the results of front-end and back-end development. Overall, the performance is satisfactory.

5.1.1. Front-end - Visualization

To visualize the traffic speed, various colors are shown to represent range of speed. For every minute, the site updates the traffic speed and changes the color of the routes if necessary. If the speed changes significantly, the color changes. The refresh time is short (< 0.1s) but the color segments are not totally fit to the routes. Since government site only provides the starting and ending nodes' coordinates of each route, the segments are straight and cannot snap to road. Goggle Map API provides a function "snap to roads" which returns points snapped to the most likely roads the vehicles travel (Figure 4). The quality is expected to be improved after implementation.

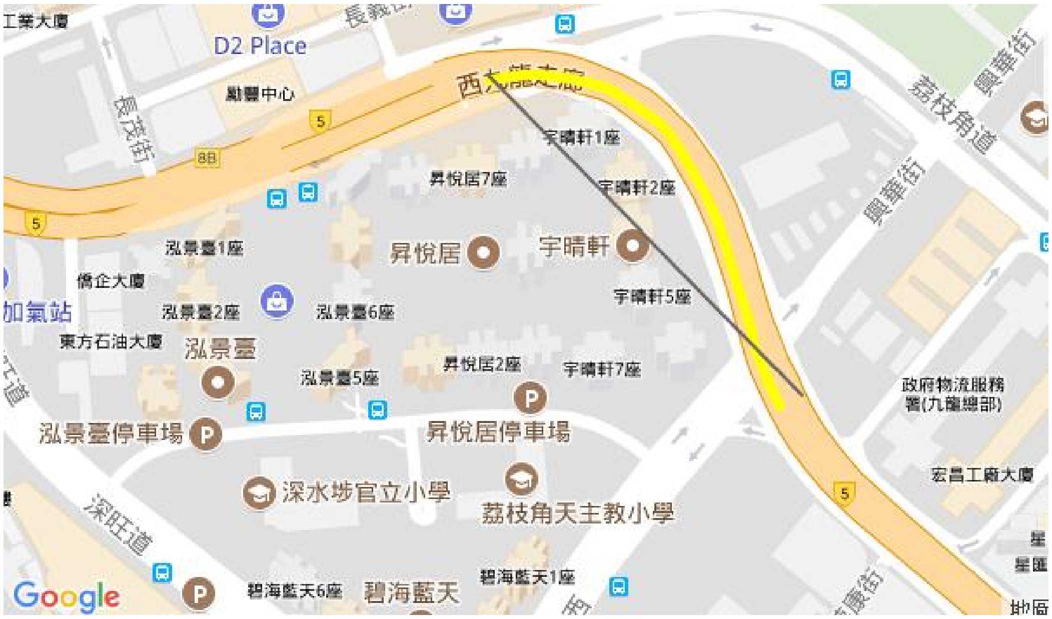


Figure 4 – Example of "Snap to road" function. The function returns the most likely roads (yellow line) according to the user input (black line).

5.1.2. Back-end –Auto data request

All the data can be captured from the government site and inserted into database successfully if the program operates normally. However, a database connection error happens irregularly and stops the program. The error cannot be identified and an alternative solution is considered. A backup program inserting the records of specific date specified by us is implemented. Although it has to be done manually, the program can minimize the adverse effect.

5.2. Forecasting

5.2.1. Forecast result

The following Table 2 shows the results of forecasting the next day using the previous 5 days record. The mean error (|Forecast - Actual|) is around 1 and the standard error is less than 1. It means that the models predict the trend well.

Data Range	Forecasting Day	MEAN	AVERAGE
2017/10/01 - 2017/10/05	2017/10/06	1.195725	0.862754
2017/10/08 - 2017/10/12	2017/10/13	0.675318	0.592515
2017/10/15 - 2017/10/19	2017/10/20	0.675034	0.595823
2017/10/22 - 2017/10/26	2017/10/27	0.726717	0.848107

Table 2 – Forecasting results

5.2.2. Model Identification process

If the model can be identified, the forecast result is acceptable. However, ARIMA model cannot fit most of the road data. In addition, the identification process is a trial and error process since the seasonality increases the complexity and the randomness is substantial. Therefore an alternative solutions are required. The current preference is using machine learning. The series may have the unknown properties and changes according to the properties. It is difficult to reveal them all and therefore the learning algorithm is required. The series data will be labeled according to the road, date and time capture in order to examine the patterns of the series data.

5.2.3. Data size selection

The model fitting depends on the size of the dataset. For instance, a model may be adequate for 5 days data but not for a week. In addition, the forecasting result is also depends day of the week but not only the last n-days data. Therefore, further research should be done to examine the relationship between the speed and day of the week.

6. Schedule

Table 3 shows the progress status, the visualization part is completed and the forecasting part is in initial stage.

Item	Expected Deadline	Completion Date
Semester 1		
Deliverables of Phrase 1 · Detailed project plan · Project website	1st October 2017	1st October, 2017
Requirement Study, technology study	20th October, 2017	18th October, 2017
Data request · Automatic process · Verification	10th November, 2017	8th November, 2017
Map implementation	17th November, 2017	10th November, 2017
Data processing · Data type conversion	8th December, 2017	12th November, 2017
Display the real-time traffic speed map	25th December, 2017	12th November, 2017
Preliminary testing	31st December, 2017	1st December, 2017
Deliverables of Phrase 2 · First presentation · Preliminary implementation · Detailed interim report	21st January 2018	21st January 2018

Semester 2		
Researching on forecasting method using machine learning (Milestone 2)	15th February 2018	TBD
Implementation the forecasting program	30th March 2018	TBD
Final review	7th April 2018	TBD
Deliverables of Phrase 3 · Finalized tested implementation · Final report · Final presentation · Project presentation	15th April 2018	TBD

Table 3 - Progress of the project

7. Conclusion

Because of the unreliable source, the performance of current map application is not satisfied. With the data from government site, my application can display the live traffic and forecast the speed for typical traffic with high accuracy. A real-time traffic speed map is completed and the application can be accessed at any time and displays the data with high accuracy. The auto data program in back-end captures all the data from government and functions normally most of the time. The performance is satisfactory and both servers and the database operate properly and no further set up are required. Nevertheless, the display quality should be improved. The error should also be identified shortly. In the forecasting part, the process time is undetermined and the machine learning is suggested. The study of machine learning is in progress and will be completed on or before mid-February and the forecasting process can be streamlined then.

8. Reference

[1] Traffic Speed Map – Transport Department, HKSAR [Internet]. [cited 2017 Nov 29]. Available from: http://www.td.gov.hk/en/transport_in_hong_kong/its/its_achievements/traffic_speed_map/index.html

[2] Visualization of Road Traffic Condition in Hong Kong – Line Wencong & Xia Fan [Internet]. [cited 2017 Sep 26]. Available from: https://zh.scribd.com/doc/245840020/visualization

[3] Data Specification for Traffic Speed Map - Transport Department, HKSAR [Internet]. [cited 2017 Nov 10]. Available from: http://static.data.gov.hk/td/traffic-speed-map/en/tsm_dataspec.pdf

[4] Explanatory Notes on Geodetic Datum in Hong Kong – Survey & Mapping Office Lands Department [Internet]. [cited 2017 Sep 26]. Available from: http://www.geodetic.gov.hk/data/pdf/explanatorynotes_c.pdf

[5] Google+ Is The Fourth Most-Used Smartphone App – Cooper Smith [Internet]. [cited 2017 Nov 29]. Available from: http://www.businessinsider.com/google-smartphone-app-popularity-2013-9#infographic

[6] Time Series Analysis With Application in R – Cryer, Jonathan D., Chan, Kung-Sik [Book]. [cited 2018 Jan 01].