

Project Report

518021911193 刘昊林

a) Introduction

After receiving the final project and analyzing the requirements mentioned in the Final-Project-CS241.pdf, I design a graphic user interface which has two components and five functions:

The two components are named **centralwidget** and **menubar**. **centralwidget** is a tabwidget and designed to implement all functions, such as selecting the parameter settings, displaying the data and plotting the trend. And it has four tabs. They are named **data**, **chart**, **road planning** and **analysis**. **menubar** is designed to be a navigation bar which lets users select the functions. And it has five parts. They are named **File**, **Choosedata**, **Chart**, **RoadPlanning** and **Analysis**.

The first function is about operations on files. It is located in the **menubar**, and its name is **File**. And it has two sub-operations, one is named **Open** and the other is named **Save**. The sub-operation **Open** is designed to select the file on the computer and open it. After selecting the file, the path of the file will be shown on the **data** tab. The sub-operation **Save** is designed to save the current file.

The second function is about importing the selected fields of user's interest based on the file user chose. It is located in the **menubar**, and its name is **Choosedata**. And it has seven sub-operations. They are named **time**, **lineID**, **stationID**, **deviceID**, **status**, **userID** and **payType**. In the initial state, they are all selected and all fields of data will be shown on the **data** tab. Once user select one of the options, the counterpart column will be hidden.

The third function is about plotting the trend of traffic inflow and outflow over time based on different data sets. It is located in the **menubar**, and its name is **Chart**. And it has two sub-operations. They are named **Based on the dataset you choose** and **Based on all of the data**. Once user chooses the first option, the program will plot the trend of traffic inflow and outflow on the **chart** tab based on the starting time, ending time, time step, station ID and the data set that user chose on the **chart** tab. Once user chooses the second option, the program will plot the trend of traffic inflow and outflow on the **chart** tab based on the starting time, ending time, time step, station ID that user chose on the **chart** tab and all of the data set.

The fourth function is about planning one or more possible routes for the user. It is located in the **menubar**, and its name is **RoadPlanning**. And it has only one sub-operation, which is named **choose a Metro_roadMap** and designed to let user select a file which is a road map. After user chooses the entering station ID and exiting station ID on the **road planning** tab, the program will plan one or more possible routes for the user on the **road planning** tab.

The fifth function is about counting and analyzing different types of data based on different data sets. It is located in the **menubar**, and its name is **Analysis**. And it has four sub-operations. They are named **lineID**, **stationID**, **deviceID** and **payType** and they all have two status, **Based on the dataset you choose** and **Based on all of the data**. This function is similar to the third one, but this function can select the field of data and the data set to analyze, and it can plot in different kinds of charts.

Figure 1 bellow is the brief mind map of my project.

Note: There is also a detailed mind map in the folde 'Final Project\project report\all figures'!!

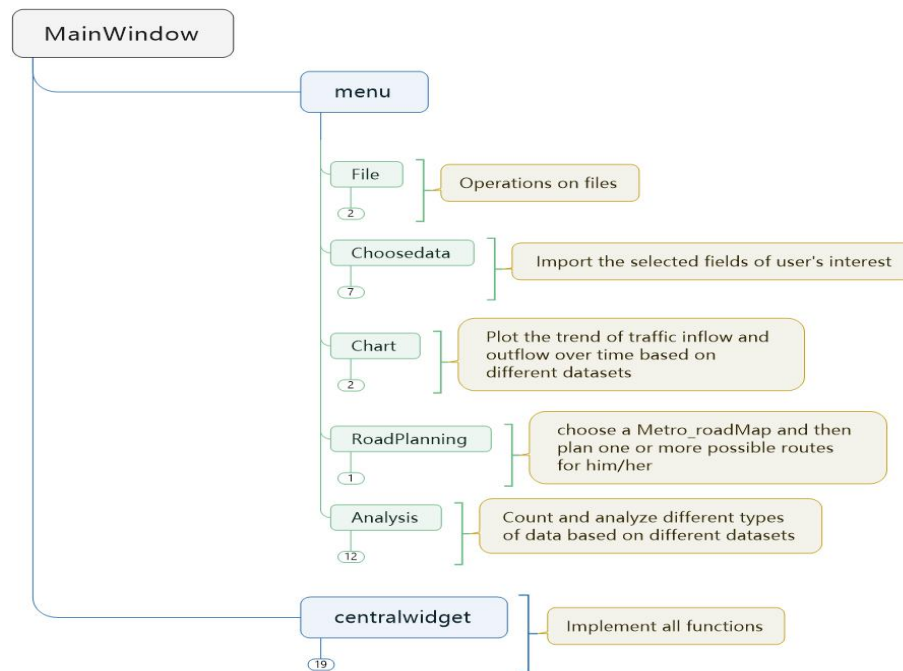


Figure 1

b) Implementation Details

First of all, I am going to introduce the interface design. In fact, most of the interface is designed by the QT designer in the QT Creator. The biggest challenge here is how to choose a right widget and it takes me a long time to figure out the function of each widget. The only problem here is that if I want to let the widgets scale with the interface, I must set the layout as grid layout.

Then, I'll focus on the implementation of five functions mentioned above.

In the first function, the **Open** and **Save** options are both turned into slots. Once they are triggered, they will send a signal, and then the functions **on_actionOpen_triggered** and **on_actionSave_triggered** will work. In the first function, it uses **QFileDialog** tool to get the path of the file which user wants to open and the path will be shown on the **data** tab. Then it uses a function **loadFile** which uses the **QTextStream** tool to read the target file line by line and at the same time using an array to store data by category and three arrays to statistic pay type condition, station ID condition and device ID condition respectively. Finally, based on the stored data and the fields selected in the second function, the program will display the data on the **data** tab. In the latter function, it uses **QFileDialog** tool to get the path of the file which user wants to save. Then it uses the **QTextStream** tool to write the stored data in the array to the target file line by line and separate data by commas.

In the second function, the **time**, **lineID**, **stationID**, **deviceID**, **status**, **userID** and **payType** options are all turned into slots and check-able. Once they are triggered, they will send a signal, and then the functions such as **on_actiontime_triggered**, **on_actionlineID_triggered** and so on will work. Because all of them are initially set as checked and each of them has a bool variable whose value is true to represent their status. Once they are triggered, the corresponding function turns the bool variable's value into false, and the counterpart column shown in the first function will be hidden on the **data** tab.

In the third function, **Based on the dataset you choose** and **Based on all of the data** options are turned into slots. Once they are triggered, they will send a signal, and then functions **on_actionBased_on_the_dataset_you_choose_triggered** and

on_actionBased_on_all_of_the_data_triggered will change the value of **type(int)** which represents the type of data set(0 means the data set is chosen by user,1 means all of the data).What's more,in the latter function,it uses **QFileDialog** tool to get the path of a station_all file.Then in both two functions,the program will set the initial starting time and ending time on the **chart** tab depends on the data set.Then,user has to select the starting time, ending time, time step and station ID on the **chart** tab.And the **Spinbox** tool,**QtimeEdit** tool and **QDateTimeEdit** tool are turned into slots.Once they are triggered,they will send a signal, then functions **on_dateTimeEdit_dateTimeChanged**,**on_dateTimeEdit_2_dateTimeChanged**, **on_spinBox_3_valueChanged** and **on_timeEdit_userTimeChanged** will change the value of **start_time**,**end_time_1**,**end_time_2**(they are all **QString**),**station_show(int)** and **count(int)**. After selecting the parameter settings,user has to click the YES button.The button is turned into slot.Once it is triggered,it will send a signal,and then function **on_pushButton_time_clicked** will work and plot the trend of traffic inflow and outflow over time based on the parameter settings by functions **plot_in_trend** and **plot_out_trend(type=0)**or **plot_in_trend_all** and **plot_out_trend_all(type=1)**.Because there is a time step,I paint both the trend of traffic inflow depends on the total number of people which isn't influenced by the time step and the the trend of traffic inflow depends on the number of people over a period of time which is influenced by the time step.The trend of traffic outflow is also the same.What's more,in order to smooth the line when only a few data points are available,I use **QSplineSeries** as the type of data series to store the data points.

In the fourth function,**choose a Metro_roadMap** is turned into slot.Once it is triggered,it will send a signal,and then function **on_actionplan_a_road_triggered** will use **QFileDialog** tool to get the path of the files store road map and total number of customers(a station_all file) in each station which user wants to open.Then the program uses the **QTextStream** tool to read the target file line by line and at the same time using an array to store total number of customers in each station.Then user has to select the the entering station ID and exiting station ID on the **road planning** tab.And these two **Spinbox** tools are turned into slots.Once it is triggered,it will send a signal,and then functions **on_spinBox_valueChanged** and **on_spinBox_2_valueChanged** will change **start(int)** and **end(int)**.After selecting the parameter settings,user has to click the YES button.The button is turned into slot.Once it is triggered,it will send a signal,and then function **on_YesButton_clicked** will use the function **findallroads** and **dfs** to find all possible routes,and plan the route with the least number of people.Finally,the program will output all the road information on the **road planning** tab.

In the fifth function,**lineID**,**stationID**,**deviceID** and **payType** are four types of data to analyze,the implementation of them is similar to the third function. They both can analyse data from the the data set chosen by user or from the whole data set(user here has to select the corresponding *_all file,*=line,station,device,pay type).But the differences are when analyzing line ID and pay type,I use the pie chart to visualize data and when analyzing station ID and device ID,I use the histogram to visualize data.

Note:There are all kinds of *_all file in the pre-processed data folder 'Final Project\data sets\preprocessed data'!!!

c) Results

Here I'm going to use some figures to show the results and the discussions will in the next part!!!

Note:There are detailed figures in the folder 'Final Project\project report\all figures'!!

	time	lineID	stationID	deviceID	status	userID	payType
1	2019-01-07 06:41:33	C	63	2952	1	C0453c92064c39bec91edd3dff...	2
2	2019-01-07 18:42:57	B	14	718	1	Cdfa39e9545b6200c680974c69...	2
3	2019-01-07 19:33:06	A	68	3158	1	C19472e2452a793f051b11c27e...	2
4	2019-01-07 21:17:24	C	49	2356	0	D025da09ee7437fc00a2176f43...	3

Figure 2 show all fields of data

	lineID	stationID	deviceID	status	userID	payType
1	C	63	2952	1	C0453c92064c39bec91edd3dff98e50b	2
2	B	14	718	1	Cdfa39e9545b6200c680974c69cd098	2
3	A	68	3158	1	C19472e2452a793f051b11c27e08b2e	2
4	C	49	2356	0	D025da09ee7437fc00a2176f4394feb3	3

Figure 3 show part of fields of data

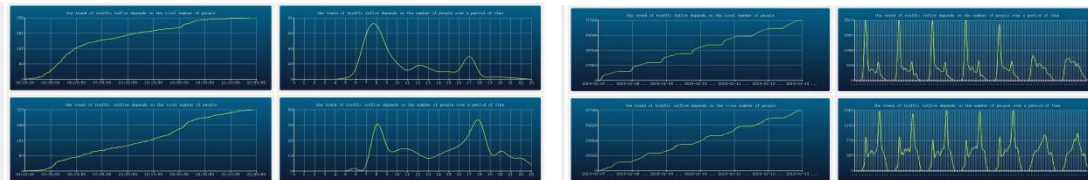


Figure 4 plot the trend on chosen data set

Figure 5 plot the trend on all data sets

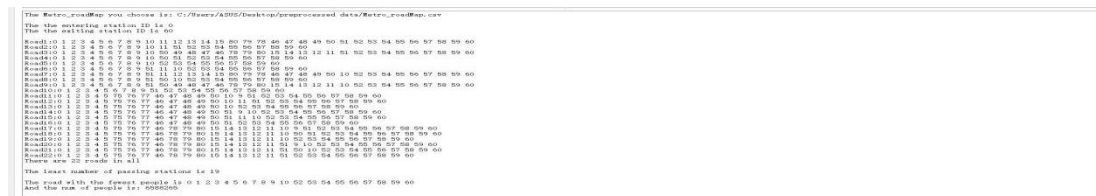


Figure 6 plan one or more possible routes for the user

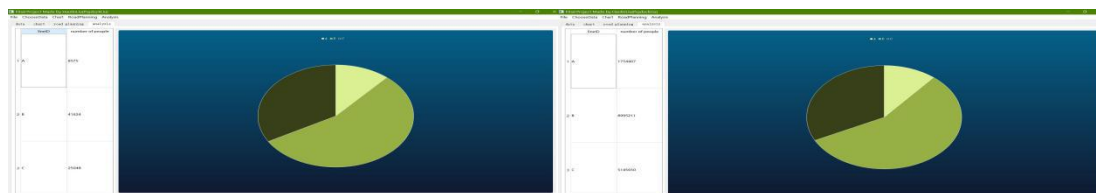


Figure 7 the line condition on chosen data set

Figure 8 the line condition on all data sets

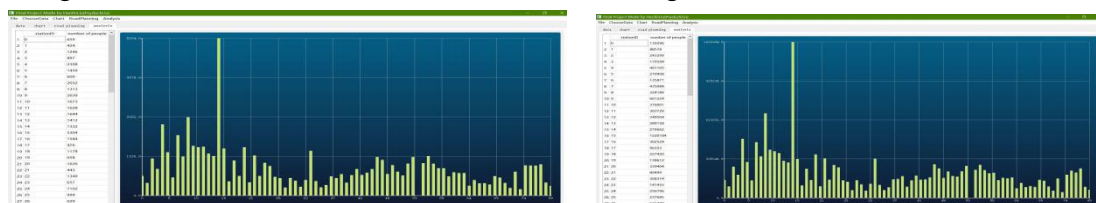


Figure 9 the station population on chosen data set Figure 10 the station population on all data sets

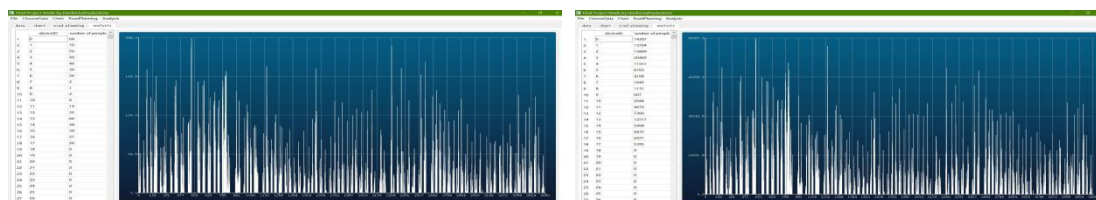


Figure 11 the device condition on chosen data set Figure 12 the device condition on all data sets

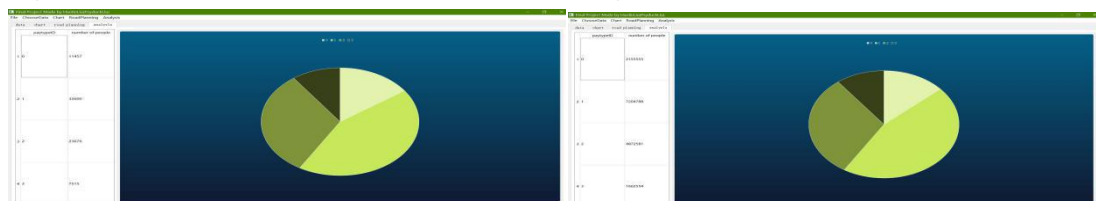


Figure13 the pay type condition on chosen data set Figure14 the pay type condition on all data sets

d) Discussions

1. the performance of my application

Each time when I want to open a file, it takes me about 5 seconds to finish it. It's acceptable to open only one file, but if I want to load all of the data sets at a time, I believe it will be a disaster. But I don't know how to use the SQLite, so I decide to find a new way to solve this problem.

I divide all analysis tasks into two parts. One is analyzing data based on the data file chosen by the user (only one file), and it spends little time. The other is analyzing data based on the data files which has been pre-processed from the whole data sets (extract some features, such as line ID, station ID, device ID and pay type.....). So even when I want to analyze the whole data sets, I also just have to load only one file. This measure really saves much time. But when I preprocess the data, it also takes a long time. Especially when I classify data according to its station and status, it spends about 5 hours doing it.

This really shocks me up! I begin to understand the importance of using database, and I make up my mind to learn it during the vacation.

2. the interesting results I have revealed from the data

① From Figure 7 and Figure 8, I can find that the usage rate among three lines is $B > C > A$. And according to the survey, I guess line B is line 1, line C is line 2 and line A is line 4 in the real world.

② From Figure 9 and Figure 10, I can find that the usage rate of station 15 is the highest. And according to the survey, this station is "杭州火车东站". This shows data tallies well with the facts.

③ From Figure 11 and Figure 12, I can find that some of the devices are frequently used, but some of them are never used. I am very puzzled with the result. I believe there must be something wrong with the data set or my code. Please teachers and TAs help me to solve this problem, thanks in advance.

④ From Figure 13 and Figure 14, I can find that the usage rate among four kinds of pay types is $1 > 2 > 0 > 3$. According to the background of the times, I guess the pay type 1 is mobile payment.

⑤ From Figure 5, I can clearly see the peak hours on each day and the difference of traffic between weekdays and weekends. So based on many figures that reflect different stations' condition, I can distinguish which stations are the places to live, which stations are the places to work and which stations are the places to relax according to the peak hours on weekdays and the number of people on weekends.

stations to live are: 0, 1, 3, 16, 22, 23, 28, 29, 32, 33, 34, 35, 36, 43, 62, 63, 64, 65, 66, 67

stations to work are: 7, 8, 9, 10, 11, 47, 50, 52, 75, 76, 77

stations to relax are: 5, 26, 29, 38, 51, 68

⑥ From all the figures, I can find that there is small difference between the chart of part of data and the chart of all of the data. This shows data tallies well with the facts. And this also confirms that Law of Large Numbers (a law in the Probability theory) is true.