ISSN No:-2456-2165

Data Analysis and Visualization of COVID-19 Epidemic based on Python

Weiyi Ma
School of Computer Science and Technology
Shandong University of Technology
Zibo City, Shandong Province, China

Dongmei Zhang*
School of Computer Science and Technology
Shandong University of Technology
Zibo City, Shandong Province, China

Abstract:- The new coronavirus pneumonia (COVID-19) that broke out at the end of 2019 was designated by the World Health Organization (WHO) as an "emergency public health event of international concern." In the process of epidemic prevention and control, big data and Internet technology have played an important role in the collection, analysis, and release of epidemic data. The purpose of the project is to implement a Python-based data analysis and visualization program for the COVID-19 epidemic. The thesis displays the epidemic situation and transmission characteristics of the existing data through a visualization scheme, establishes a dynamic model of infectious diseases, evaluates the prevention and control measures of the epidemic situation, and makes recommendations and early warnings. In addition, to a certain extent, it can predict the trend of epidemic diseases and provide reference for epidemic prevention and control decisions and public behavior.

Keywords:- *Novel coronavirus pneumonia; COVID-19; Python; data analysis; data visualization.*

I. INTRODUCTION

A. Subject Background and Significance

a) Background

Pneumonia caused by a new type of coronavirus was discovered in Wuhan, Hubei in December 2019, and it is showing a trend of rapid spread. On February 7, 2020, the National Health Commission (PRC) named it "New Coronavirus Pneumonia", or "New Coronary Pneumonia" for short. On February 11, 2020, the International Commission for Classification of Viruses (ICTV) named the virus Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2); on the same day, the World Health Organization (WHO) The disease it caused was named Coronavirus Disease 2019 (COVID-19). In 2020, the new crown pneumonia epidemic has broken out one after another around the world, which has extremely serious impacts on the global economy and society, and has caused great troubles to human health and life [1].

In various aspects of epidemic prevention and control, apart from the effective measure of isolation, scientific popularization of daily epidemic prevention and control knowledge, timely release of epidemic transmission and infection data, etc., can enable the public to understand the epidemic in a timely manner, take reasonable response measures, and avoid panic. The occurrence of bad consequences caused by spread. The collection, sorting, analysis, and visualization of data can be completed well with the help of big data technology. This topic is based on the Python-based data analysis and visualization program of the COVID-19 epidemic, and is proposed based on this background.

b) Meaning

The new crown epidemic spread rapidly to more than 200 countries and regions around the world within a few months, and as of the beginning of June, there have been more than 6 million patients. China has achieved full control of the epidemic in May, and a number of response measures are worthy of promotion. In the information age, with the help of big data and artificial intelligence technology, it is possible to quickly establish an effective system and mechanism for responding to public health emergencies. Its intuitive and effective data analysis methods and artificial intelligence visualization methods have played a pivotal role.

The topic is based on prediction models such as SEIR, taking the data of COVID-19 epidemic in Hubei Province as an example, preliminary analysis of the general law of COVID-19 epidemic, and a prediction analysis. From the perspective of the prevention and control process of the new crown epidemic, studying the occurrence, development, and evolution of the epidemic from a macro perspective, and predicting and analyzing it based on big data, are of great significance to the strategic decision-making of large-scale prevention and control of infectious diseases and maintaining social order and stability.

B. Current Research Status at Home and Abroad

In the early stage of the outbreak, many scholars tried to study and analyze the development trend of the new crown epidemic through the infectious disease dynamic model.

In January of this year, many scholars predicted the epidemic. Wu, a scholar from Hong Kong, China, used the number of people infected before January 28 to calculate the trend of the epidemic in Wuhan. They predict that the number of infections on January 25 will exceed 6,000. Professor Shen and others from Xi'an Jiaotong University estimated that the number of SARS-CoV-2 infections will not exceed SARS-CoV-2 based on the existing epidemiological data and infectious disease dynamic models, and with reference to SARS and other coronaviruses. 20,000 people, but this is lower than the epidemic data released on February 7, which obviously underestimates the infectiousness of the new coronavirus. Professor Xiao from Xi'an Jiaotong University established an infectious disease dynamics model based on domestic and foreign research on the transmission mechanism of the new coronavirus, based on strict tracking and isolation measures. The risk of transmission of the new coronavirus pneumonia was predicted and analyzed, and the mission will reach the peak of the epidemic in February. However, the current epidemic situation has exceeded the predicted result. In February of this year, some researchers tried to use models such as the SEIR model and c-SEIR model to infer the "turning point" of the epidemic, but in these models, they did not fully consider factors such as control measures and intensive treatment that cannot be ignored in practice [2].

C. *Introduction To Key Technologies*

➢ Introduction to Data Visualization Technology

- Pandas: Pandas is a Python data analysis software package developed by AQR Capital Management in April 2008 and released as an open source at the end of 2009. Pandas was originally developed as a financial data analysis tool, so it provides good support for time series analysis. Pandas is a powerful tool set for analyzing structured data. It is based on NumPy, which is used for data mining and data analysis. It also provides data cleaning functions.
- ECharts and Pyecharts: ECharts is the abbreviation of Enterprise Charts. Enterprise-level data charts are a pure Javascript chart library that can run smoothly on PC and mobile devices. It is compatible with most current browsers and provides intuitive, vivid, interactive and highly interactive. Personalized customized data visualization chart. Innovative drag-and-drop calculations, data views, range roaming and other features greatly enhance the user experience and enable users to mine and integrate data. Pyecharts is a class library for generating Echarts charts. Pyecharts is for docking with Python to facilitate the direct use of data to generate graphs in Python.

c) Infectious Disease Prediction Model

- SEIR model: The classic SEIR model divides the population into susceptible (S), infected (I), lurking (exposed, E), and recovered (R). The model also assumes that all individuals in the population have the probability of being infected. When the infected individuals recover, they will produce antibodies, that is, the recovered population R will not be infected again.
- Improved SEIR model: Due to the isolation measures for the prevention and control of infectious diseases, we can group the population in the model to add the susceptible person Sq, the latent person Eq and the infected person Iq. Because quarantined and infected people will be sent to designated hospitals for quarantine treatment, this part of the population will be transformed into hospitalized patients H in this model [3]. Therefore, in the improved model, S, I, and E respectively refer to the susceptible, infected, and latent people missed under the isolation measures. Isolated susceptible persons can be transformed into susceptible persons again after being released from quarantine, while infected persons and latent persons have different degrees of ability to infect susceptible persons, which can transform them into latent persons.

d) Data Sources

This project uses web crawlers, and the data comes from the real-time epidemic website of Dingxiangyuan, the real-time epidemic website of Tencent and the hot search website of Baidu epidemic.

II. SYSTEM DESIGN

A. *System Overall Architecture Design*

For the convenience of user operation, the system uses B/S architecture as the basic system architecture. Due to the need to consider data security and system stability, the application is deployed separately during deployment.

As shown in Figure 1, the user accesses the browser, and sends a network request to the back end of the system, and then the back end receives this request and visits the view layer. The view layer obtains the data in the database by accessing the database model interface, and then uploads and returns layer by layer. The view layer places the obtained return data into variables in the template, and finally displays the page to the user.

The data analysis and visualization system is divided into three parts: data acquisition module, data analysis module and analysis result display module. The platform uses Python 3 for development, the Web development framework uses Flask, and the database uses MySQL. The analysis of structured data uses Pandas, the word segmentation processing of unstructured data uses Jieba, and the visualization uses Echarts. The integrated development environment uses Jupter Notebook and PyCharm.

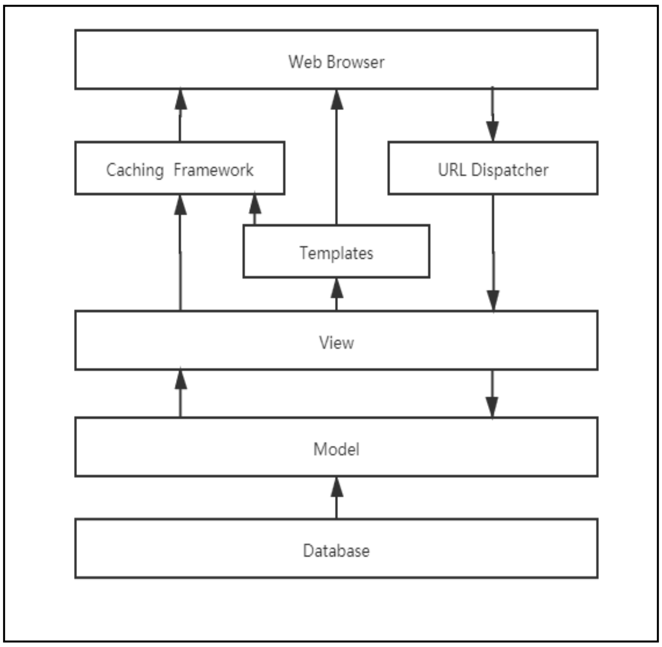


Fig. 1 : System Architecture Model

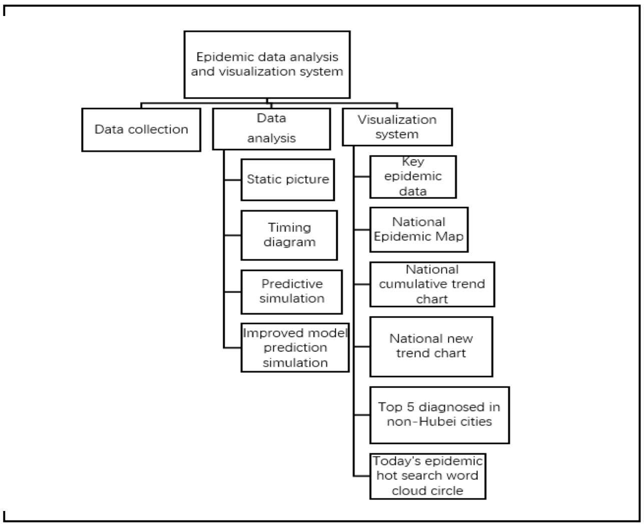
B. *System Module Design*

According to the system requirements analysis content and system architecture design, the system can be divided into three modules: data acquisition, data analysis and visualization system. The system module structure diagram is shown in Figure 2.

- Data collection module: Use python crawler technology to crawl data from epidemic data websites, use selenium web automation tools to simulate the operation of Chrome browser web pages, and use requests to obtain web page information. Either manually run the crawler file to obtain the specified data, or deploy it to the server environment to run automatically.

System architecture model.

- Data analysis module: use Pandas for data reading and grouping aggregation calculations, use Pandas+Seaborn for visual display and simple time series analysis, use Folium for geographic-based data visualization, use Seaborn to draw heat maps to display hotspots, and use SEIR Wait for the epidemic model to predict the epidemic situation, and finally form a visual chart, such as a geographic map, a trend chart, and an epidemic forecast simulation map [4].


Epidemic data analysis and visualization system
Data collection
Data analysis
Visualization system
Static picture
Timing diagram
Predictive simulation
Improved model prediction simulation
Key epidemic data
National Epidemic Map
National cumulative trend chart
National new trend chart
Top 5 diagnosed in non-Hubei cities
Today's epidemic hot search word cloud circle

- Visualization system module: A large data visualization screen based on the Flask framework and Echarts technology provides users with an intuitive understanding of domestic epidemic information, displaying key epidemic data, national epidemic maps, epidemic trend graphs, non-Hubei city confirmed rankings and epidemic hot searches Word cloud illustration.

III. SYSTEM IMPLEMENTATION

A. The Realization Of The Data Acquisition Module

Use the selenium web automation tool to simulate the operation of the Chrome browser web page, and use requests to obtain web page information. Either manually run the crawler file to obtain the specified data, or deploy it to the server environment to run automatically. After the module runs, update detailed epidemic data, epidemic history data, or epidemic hot search data according to the obtained parameters.

The core code is as follows:

```
def get_baidu_hot():

option = ChromeOptions()

option.add_argument("--headless")

url='https://voice.baidu.com/act/virussearch/virussearch?from=osari_map&tab=0&infomore=1'

brower = Chrome(options = option)

brower.get(url)

but = brower.find_element_by_css_selector('#ptab-0 > div > div.VirusHot_1-5-6_32AY4F.VirusHot_1-5-6_2RnRvg > section > div')

but.click()

time.sleep(1)

c = brower.find_elements_by_xpath('//*[@id="ptab-0"]/div/div[1]/section/a/div/span[2]')

context = [i.text for i in c]

print(context)

return context
```

B. The Realization Of The Data Analysis Module

This module realizes the distribution of domestic and foreign epidemic data, analyzes the epidemic distribution, spread trend, development trend and epidemic forecast simulation. The module interface is shown in the following figure.

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

a) Comparison of epidemics at home and abroad

A comparison of domestic and foreign advanced studies from the four aspects of current diagnoses, cumulative diagnoses, cured numbers and deaths. Using a pie chart, you can show the proportion of each part to the whole.

The core code is as follows:

```
def new_label_opts():

return opts.LabelOpts(formatter=JsCode(fn), position="center")

pie = (Pie(init_opts=opts.InitOpts(theme='dark', width='1000px'))

.add(
" Diagnosed on the same day",

 [(x, y) for x, y in oversea_data['currentConfirmedCount'].items()],

center=["30%", "30%"],

radius=[60, 90],

label_opts=new_label_opts(),
)

.add(

" Cumulative diagnosis",

[(x, y) for x, y in oversea_data['confirmedCount'].items()],
center=["70%", "30%"],

radius=[60, 90],

label_opts=new_label_opts(),

)

.set_global_opts(

title_opts=opts.TitleOpts(title=" Comparison of epidemic data at home and abroad ",

subtitle=" Update time：{}".format(update_date)),

legend_opts=opts.LegendOpts(is_show=True),)

.set_series_opts(

tooltip_opts=opts.TooltipOpts(

trigger="item", formatter="{a} <br/>{b}: {c} ({d}%)"

))
```

b) National Epidemic Map

The national epidemic map can clearly reflect the distribution of the epidemic. The more cumulatively confirmed cases, the heavier the regional color.

The core code is as follows:

```
_map = (

Map(init_opts=opts.InitOpts(theme='dark',width='1000px'))

add("Cumulative confirmed number", cofirm, "china", is_map_symbol_show=False, is_roam=False)
.set_series_opts(label_opts=opts.LabelOpts(is_show=True))

.set_global_opts(

title_opts=opts.TitleOpts(title=" National Epidemic Map of the Novel Coronavirus ",

subtitle="Update time：{}".format(update_date)),

legend_opts=opts.LegendOpts(is_show=False),

visualmap_opts=opts.VisualMapOpts(is_show=True, max_=1000,

is_piecewise=False,

range_color=['#FFFFE0', '#FFA07A', '#CD5C5C', '#8B0000'])
)
)
```

c) Trend map of cumulative diagnoses nationwide

The nationwide cumulative diagnosis trend map can intuitively reflect the development trend of the epidemic, and at the same time compare the whole country, Hubei and Wuhan together to reflect the relationship between them.

The core code is as follows:

```
for key_, value_ in data_type.items():

line = (Line(init_opts=opts.InitOpts(theme='dark', width='1000px'))

.add_xaxis([day.strftime('%Y-%m-%d') for day in time_range])

.add_yaxis("The entire country", area_data('The entire country', value_), is_smooth=True,

areastyle_opts=opts.AreaStyleOpts(opacity=0.5,

.add_yaxis("Hubei", area_data('Hubei', value_), is_smooth=True,

areastyle_opts=opts.AreaStyleOpts(opacity=0))

.add_yaxis("Wuhan", area_data('Wuhan', value_), is_smooth=True,

areastyle_opts=opts.AreaStyleOpts(opacity=0))

.set_series_opts(label_opts=opts.LabelOpts(is_show=False))

.set_global_opts(

title_opts=opts.TitleOpts(title="The trend graph{}of the entire country ".format(key_),

subtitle="Update time：{}".format(update_date)),

))
```

d) National epidemic heat map

The national epidemic heat map can show the severity of the epidemic at a certain time node in various regions of the country over time, reflecting the spread and development of the epidemic.

The core code is as follows:

```
for day in time_range:

geo = (

Geo(init_opts=opts.InitOpts(theme='dark'))

.add_schema(maptype="china", zoom=1)

.add("Current confirmed number",
[(key_, value_['currentConfirmedCount']) for key_, value_ in format_data[day].items()

if key_ in pyecharts.datasets.COORDINATES.keys() and value_['is_city'] == 1],

type_='heatmap',
    symbol_size=3,
    progressive=50)

    .set_series_opts(label_opts=opts.LabelOpts(is_show=False))

    .set_global_opts(

    title_opts=opts.TitleOpts(title=" New Coronavirus National Epidemic Heat Map",

    subtitle="Update time：{}".format(update_date)),

    legend_opts=opts.LegendOpts(is_show=False), range_color=['blue', 'green', 'green', 'yellow', 'red']),
```

e) SEIR model prediction simulation

The SEIR model predicts the simulation map based on the SARS epidemic, setting γ=0.0821, λ=0.2586, the initial susceptible number is 10 million, the initial infection is 10, the initial migrant is 5, and the total number of people in the city is N=1 e 7+10+5, bring it into the model to get the result.

The core code is as follows:

```
# initial infective people

i[0] = 10.0 / N

s[0] = 1e7 / N

e[0] = 40.0 / N

for t in range(T-1):

    s[t + 1] = s[t] - lamda * s[t] * i[t]

    e[t + 1] = e[t] + lamda * s[t] * i[t] - sigma * e[t]

    i[t + 1] = i[t] + sigma * e[t] - gamma * i[t]

    r[t + 1] = r[t] + gamma * i[t]

fig, ax = plt.subplots(figsize=(10,6))

ax.plot(s, c='b', lw=2, label='S')

ax.plot(e, c='orange', lw=2, label='E')

ax.plot(i, c='r', lw=2, label='I')

ax.plot(r, c='g', lw=2, label='R')

ax.set_xlabel('Day',fontsize=20)

ax.set_ylabel('Infective Ratio', fontsize=20)

ax.grid(1)
plt.xticks(fontsize=20)

plt.yticks(fontsize=20)

plt.legend();
```

f) Improved SEIR model

The improved SEIR model mainly analyzes the Hubei region, uses the public data of Hubei Province as the parameter basis, considers the infectivity of lurkers and various isolation prevention and control measures, and analyzes the impact of each factor on the development of the epidemic.

The core code is as follows:

```
[S1, S2] = deal(59170000);

[E1, E2] = deal(4007);

[I1, I2] = deal(786);

[Sq1, Sq2] = deal(2776);

[Eq1, Eq2] = deal(400);

[H1, H2] = deal(1186);

[R1, R2] = deal(31);

T=1:150;

for idx =1:length(T)-1

S1(idx+1)=S1(idx)-(rho*c*beta+rho*c*q*(1-beta))*S1(idx)*(I1(idx)+theta1*E1(idx))+lambda*Sq1(idx);

E1(idx+1)=E1(idx)+rho*c*beta*(1-q)*S1(idx)*(I1(idx)+theta1*E1(idx))-sigma*E1(idx);

I1(idx+1)=I1(idx)+sigma*E1(idx)-(deltaI+alpha+gammaI)*I1(idx);

Sq1(idx+1)=Sq1(idx)+rho*c*q*(1-beta)*S1(idx)*(I1(idx)+theta1*E1(idx))-lambda*Sq1(idx);

Eq1(idx+1)=Eq1(idx)+rho*c*beta*q*S1(idx)*(I1(idx)+theta1*E1(idx))-deltaq*Eq1(idx);

H1(idx+1)=H1(idx)+deltaI*I1(idx)+deltaq+Eq1(idx)-(alpha+gammaH)*H1(idx);

R1(idx+1)=R1(idx)+gammaI*I1(idx)+gammaH*H1(idx);
End
```

C. *Implementation Of The Visualization System*

This module realizes a visual display of the domestic epidemic situation, displaying key epidemic data, a national epidemic map, an epidemic trend graph, a ranking of confirmed cases in non-Hubei cities, and an epidemic word cloud map.

The core code is as follows:

```
app = Flask(__name__)
@app.route("/r2")

def get_r2_data():

data = utils.get_r2_data()
```

```
((' Police fight on the front line to fight the epidemic for 16 days and sacrifice 1037364',), (' Sichuan sends two more medical teams 1537382',)

d = []

for i in data:

k = i[0].rstrip(string.digits)

v = i[0][len(k):]

ks = extract_tags(k)

for j in ks:

if not j.isdigit():

d.append({"name": j, "value": v})

return jsonify({"kws": d})
```

IV. CONCLUSION

This paper mainly discusses the analysis, design and implementation of each module of the new crown epidemic data analysis and visualization system. This topic conducted a preliminary analysis of the needs of the population concerned about the epidemic, and subdivided the modules. After that, a detailed design was carried out, the functional goals of the system were extracted, and the codes for each function were written, and the following functions were realized, including data collection, data analysis, and visual display. Finally, a system test was carried out to make the system more scientific and rigorous[5].

During the realization of this subject, the data analysis and visualization system is divided into three parts: data acquisition module, data analysis module and analysis result display module. The platform uses Python 3 for development, the Web development framework uses Flask, and the database uses MySQL. The analysis of structured data uses Pandas, the word segmentation processing of unstructured data uses Jieba, and the visualization uses Echarts. The integrated development environment uses Jupter Notebook and PyCharm.

Tests show that the system has completed most of the functions refined in the system's demand analysis. Due to time reasons, this system still has various deficiencies. For example, some functions have conflicts due to later modification, and the details of the interface have yet to be dealt with, settings Some settings on the page involve a wide range of areas and need to be modified slowly. I hope that through future learning, I will spare time to continuously improve it, so that the above problems can be solved more perfectly.

V. ACKNOWLEDGMENT

Dongmei Zhang is the corresponding author.

REFERENCES

[1.]Chenghu Zhou ac†, Fenzhen Su a c e †, C T P A , et al. COVID-19: Challenges to GIS with Big Data[J]. Geography and Sustainability, 2020, 1(1):77-87.

[2.]Mengistie T T. COVID-19 Outbreak Data Analysis and Prediction Modeling Using Data Mining Technique[J]. International Journal of Computer (IJC), 2020, Volume 38(No 1):pp 37-60.

[3.]Al-Rousan N,Al-Najjar H. Data Analysis of Coronavirus CoVID-19 Epidemic in South Korea Based on Recovered and Death Cases[J]. Journal of Medical Virology, 2020.

[4.]de León, Ugo Avila-Ponce, Pérez, ngel G. C, Avila-Vales E . A data driven analysis and forecast of an SEIARD epidemic model for COVID-19 in Mexico[J]. 2020.

[5.]Meiling Z , Kun W , Xiao L , et al. Epitranscriptome analysis of COVID-19 prevention and control[J]. Chinese Journal of Medical ence Research Management, 2020, 33(00):E002-E002.