Design and Implementation of Logistics Model 2019-NCOV Epidemic Development Simulation System in the Context of Big Data

Dong-Xu Liu[1*], Feng Qin[1, 2], Wan-Sheng Wang[1], Xiu-Ru Li[1], Lu-Lu Zhang[1], Lin Jiang[1], Ying-Ming Wang[1], Dan-Dan Zuo[1], Yi-Ran Zheng[1]

[1] Tencent Cloud Big Data Institute, Ma 'Anshan University,Ma 'anshan city, China.
[2]School of Computer science, Anhui University of Technology, China.

* Corresponding author. Tel.: 18726048223; email: Ldx1758465@163.com

Manuscript submitted March 14, 2022; accepted May 10, 2022.
doi: 10.17706/jsw.17.182-191

Abstract: The outbreak of COVID-19 in December 2019 sent tens of thousands of people into panic; For sudden outbreak, due to the variability and immune virus, in view of the research and development will be to contain virus vaccine, so the country should not only possess strong vaccine research and development team and scientific research ability, and more need a kind of epidemic development simulation platform, the simulation of the real effective epidemic trends, provide high quality data for prevention and research department reference value; In this paper, the minimum deviation is sought through Logistics prediction model to improve the reliability and authenticity of epidemic simulation and prediction data. Real-time communication is completed by combining big data technology Spark training model and Kafka, and the simple and intuitive H5 realistic visual interface is adopted.

Key words: Logistics model. 2019-NCOV epidemic, simulation system, big data.

1. Introduction

1.1. Big Data

Over the past half century, with the full integration of computer technology into social life, the information explosion has accumulated to a point where it is beginning to cause change [1]-[3]. Not only is it flooding the world with more information than ever before, it is also growing faster. The Internet (social networking, search, e-commerce), mobile Internet (Weibo), Internet of Things (sensors, smart Earth), Internet of vehicles, GPS, medical imaging, safety and health, finance (banking, stock market, insurance), telecommunications (call, SMS) are all generating data like crazy, as shown in Fig. 1.

The world sends 2.9 million e-mails every second. Reading one article a minute is enough for a person to read around the clock for 5.5 years. There are 28,800 hours of video uploaded to Youtu every day, enough for one person to watch around the clock for 3.3 years; There are 50 million tweets on Twitter every day. If you read one message for 10 seconds, that's enough for a person to read it day and night for 16 years. 6.3 million orders are made on Amazon every day; Every month, Internet users spend 700 billion minutes on Facebook, and 1.3 exabytes of data are sent and received by mobile Internet users. Google handles 24 petabytes of data a day; According to IDC [4]-[5], the amount of data generated by humans is increasing exponentially, doubling roughly every two years, and this rate will continue until 2020, meaning that humans have generated as much data in the last two years as they did in the past. As the amount of data continues to increase, the data structure becomes increasingly complex, and the emergence of a large number of new data sources leads to the explosive growth of unstructured and semi-structured data, as shown in Fig. 2 and Fig. 3.

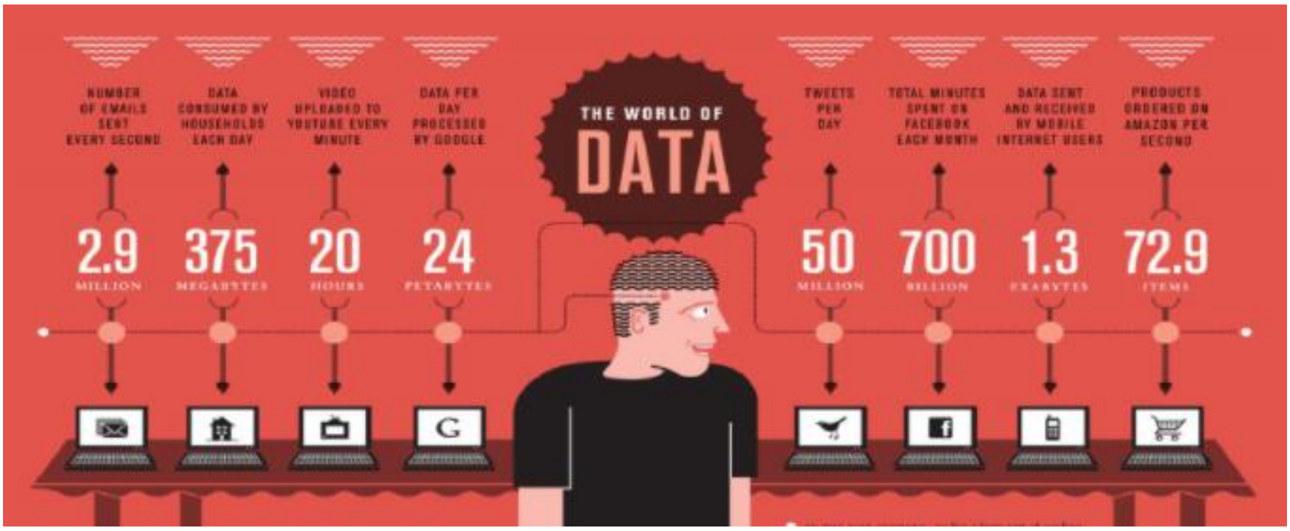
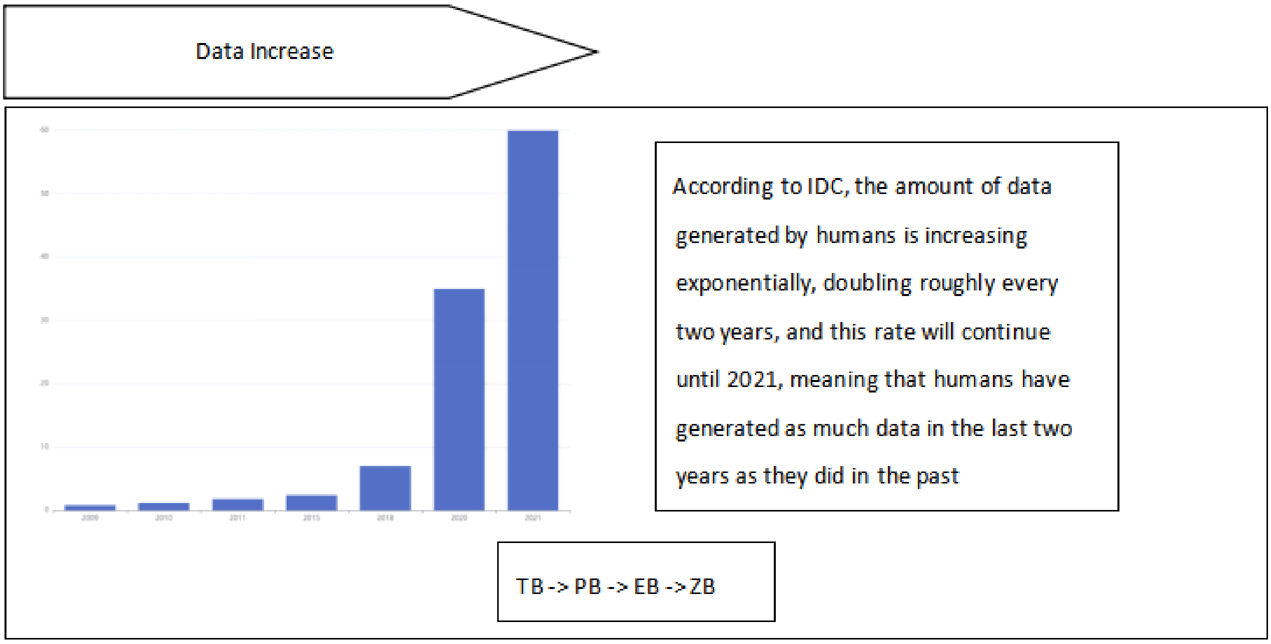

THE WORLD OF
DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS
DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES
TWEETS PER DAY
50 MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS
G

Fig. 1. The world of data.

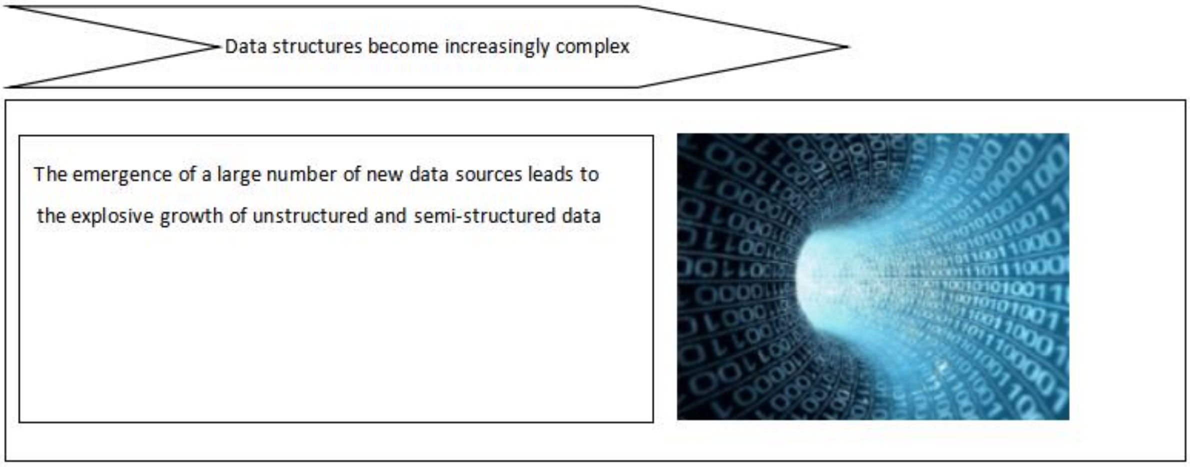


According to IDC, the amount of data generated by humans is increasing exponentially, doubling roughly every two years, and this rate will continue until 2021, meaning that humans have generated as much data in the last two years as they did in the past

Fig. 2. Data increase.

1.2. Logistics Model

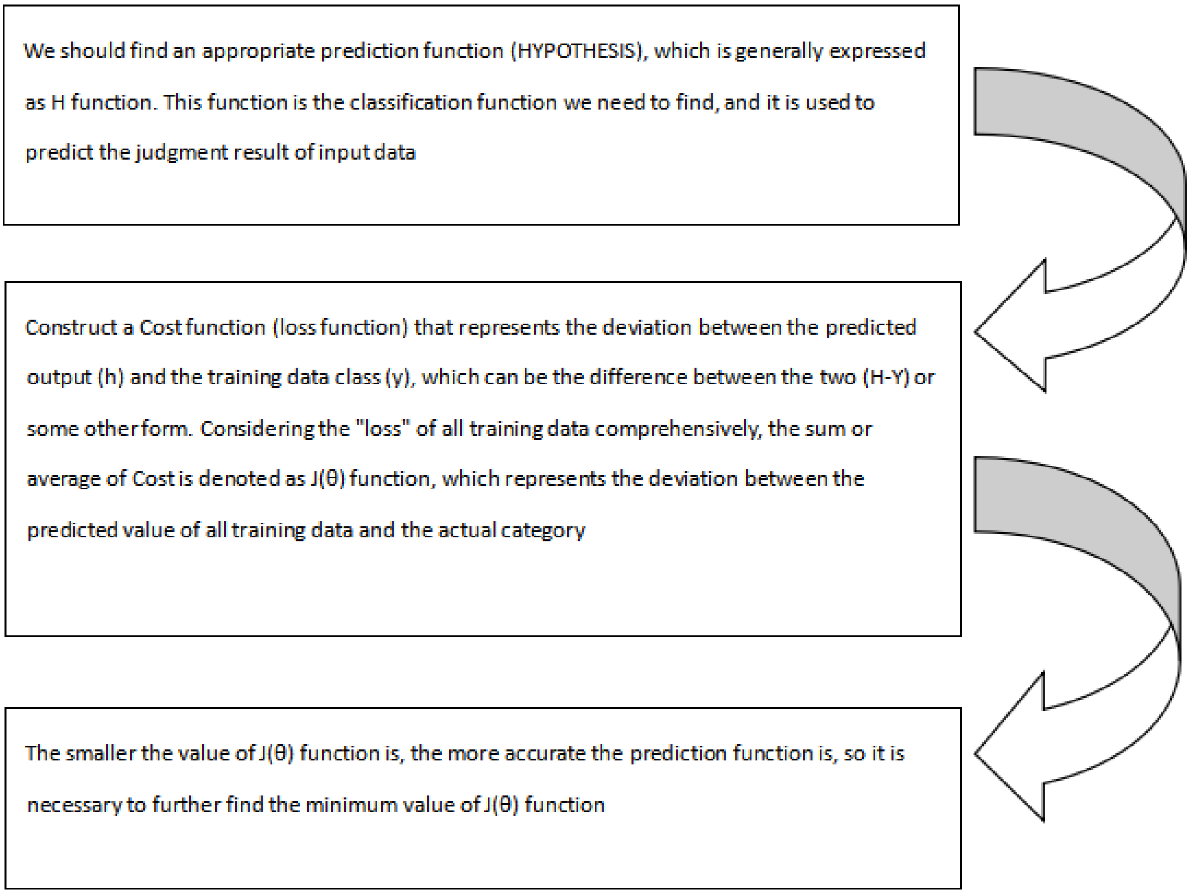
Logistic regression [6], also known as Logistic regression analysis, is mainly used in epidemiology. It is commonly used to explore the risk factors of a disease and predict the probability of the occurrence of a disease according to the risk factors, etc. For example, to discuss the risk factors of gastric cancer, two groups of people can be selected, one group is gastric cancer group, the other group is non-gastric cancer group, the two groups of people must have different physical signs and lifestyle, etc. The dependent variable here is -- whether gastric cancer, namely "yes" or "no", is two classification variables, independent variables can include a lot of, such as age, gender, eating habits, helicobacter pylori infection, etc. The independent variable can be continuous or discrete. Through logistic regression analysis, we can roughly understand what are the risk factors of gastric cancer [7]-[9].



Data structures become increasingly complex

The emergence of a large number of new data sources leads to the explosive growth of unstructured and semi-structured data

Fig. 3. Data structures become increasingly complex.



We should find an appropriate prediction function (HYPOTHESIS), which is generally expressed as H function. This function is the classification function we need to find, and it is used to predict the judgment result of input data

Construct a Cost function (loss function) that represents the deviation between the predicted output (h) and the training data class (y), which can be the difference between the two (H-Y) or some other form. Considering the "loss" of all training data comprehensively, the sum or average of Cost is denoted as J(θ) function, which represents the deviation between the predicted value of all training data and the actual category

The smaller the value of J(θ) function is, the more accurate the prediction function is, so it is necessary to further find the minimum value of J(θ) function

Fig. 4. Logistics forecasting model method flow.

2. Logistics Forecasting Model Method Submission

The first step is to find an appropriate prediction function (HYPOTHESIS), which is generally expressed as H function. This function is the classification function we need to find, and it is used to predict the judgment result of input data. This process is critical, requiring some understanding or analysis of the data, and knowing or guessing the "rough" form of the prediction function, such as whether it is linear or non-linear. The second step is to construct a Cost function (loss function), which represents the deviation between the predicted output (h) and the training data category (y), which can be the difference between the two (H-Y) or some other form. Considering the "loss" of all training data comprehensively, the sum or average of Cost is denoted as J(θ) function, which represents the deviation between the predicted value of all training data and the actual category. Third, the smaller the value of J(θ) function is, the more accurate the prediction function is (that is, the more accurate the h function is) [10], as shown in Fig. 3.

3. The Reasoning of Logistics Prediction Algorithm

1) Set the function, $\hat{\mathrm{y}}=\mathrm{f}(\mathrm{x})$.

2) In logistic regression, the value of y obtained is essentially a probability value p, $\hat{\mathrm{p}}=\mathrm{f}(\mathrm{x})$, After the probability value p is obtained, classification is carried out according to the probability value $\hat{y}=\begin{Bmatrix}1, \hat{p}\geq 0.5\\ 0, \hat{p}\leq 0.5\end{Bmatrix}$,1 and 0 May have different meanings in different situations. For example, 0 May represent patients with malignant tumors and 1 patients with benign tumors. In linear regression, we use $\hat{\mathrm{y}}=\mathrm{f}(\mathrm{x})\Rightarrow\hat{y}=\theta^T\cdot x_b$, because θ 0 exists, x is represented by a small Xb, that is, for every sample we take, we have to add another 1 in front of it. This 1 multiplied by θ 0 is the intercept, but anyway, in this case, the range of y is -infinity, +infinity, For probability, there is a limit, its range is [0,1]; We pass the result of linear regression as an eigenvalue into a new function, which is transformed into a value with a range between [0,1], $\hat{p}=\sigma\left(\theta^T\cdot x_b\right)$,Logistics function $\sigma\left(t\right)=\frac{1}{1+e^{-t}}$; Using the Logistics function, $\sigma\left(\theta^T\cdot x_b\right)=\frac{1}{1+e^{-\theta^T\cdot x_b}}$, $\hat{y}=\begin{Bmatrix}1, \hat{p}\geq 0.5\\ 0, \hat{p}\leq 0.5\end{Bmatrix}$.

3) Loss function of logistic regression, I defined a loss function that looks like this $\cos t=\begin{cases}-\log(\hat{p}), y=1\\ -\log(1-\hat{p}), y=0\end{cases}$, To simplify it, $\cos t=-y\log(\hat{p})-(1-y)\log(1-\hat{p})$, If I average the losses, $J(\theta)=-\frac{1}{m}\sum_{i=1}^{m}y^{(i)}\log(\hat{p}^{(i)})+(1-y^{(i)})\log(1-\hat{p}^{(i)})$, $\hat{p}^{(i)}=\sigma(X_b^i\theta)=\frac{1}{1+e^{-X_B^I\theta}}$, the two equations are integrated, $J(\theta)=-\frac{1}{m}\sum_{i=1}^{m}y^{(i)}\log(\sigma(X_b^i\theta))+(1-y^{(i)})\log(1-\sigma(X_b^i\theta))$.

4) First, find the formula for the gradient of J(θ), $\nabla J(\theta)=\begin{pmatrix}\frac{\partial J(\theta)}{\partial\theta_0}\\ \frac{\partial J(\theta)}{\partial\theta_1}\\ \cdots\cdots\\ \frac{\partial J(\theta)}{\partial\theta_m}\end{pmatrix}$,take the derivative of Logistics function, $\sigma\left(t\right)^{'}=(1+e^{-t})^{-2}\cdot e^{-t}$, Take the derivative of log sigma, $(\log\sigma(t))^{'}=1-\sigma(t)$,Thus, the derivative of the

first half is: $d(y^{(i)}\log\sigma(X_b^{(i)}\theta))\frac{1}{d\theta_j} = y^{(i)}(1-\sigma(X_b^{(i)}\theta))\cdot X_j^{(i)}$, where y(I) is a constant, and then take the derivative of the second half, $(\log(1-\sigma(t)))^{'} = \frac{1}{1-\sigma(t)}\cdot(-1)\cdot\sigma(t)^{'} = -\frac{1}{1-\sigma(t)}\cdot(1+e^{-t})^{-2}\cdot e^{-t}$, Simplifying the results, $\frac{d((1-y^{(i)})\log(1-\sigma(X_b^{(i)}\theta)))}{d\theta_y} = (1-y^{(i)})\cdot(-\sigma(X_b^{(i)}\theta))\cdot X_j^{(i)}$, add the front and the back, $y^{(i)}X_j^{(i)} - y^{(i)}\sigma(X_b^{(i)}\theta)\cdot X_j^{(i)} - \sigma(X_b^{(i)}\theta)\cdot X_j^{(i)} + y^{(i)}\sigma(X_b^{(i)}\theta)\cdot X_j^{(i)} = (y^{(i)} - \sigma(X_b^{(i)}\theta))\cdot X_j^{(i)}$,For a quick, $\frac{J(\theta)}{\theta_j} = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})\cdot X_j^{(i)}$, So you get, $\frac{J(\theta)}{\theta_j} = \frac{1}{m}\sum_{i=1}^{m}(\sigma(X_b^{(i)}\theta) - y^{(i)})\cdot X_j^{(i)}$,

$$\nabla J(\theta) = \begin{pmatrix} \frac{\partial J(\theta)}{\partial\theta_0} \\ \frac{\partial J(\theta)}{\partial\theta_1} \\ \cdots\cdots \\ \frac{\partial J(\theta)}{\partial\theta_m} \end{pmatrix} = \frac{1}{m}\begin{pmatrix} \sum_i^m(\sigma(X_b^{(i)}\theta) - y^{(i)}) \\ \sum_i^m(\sigma(X_b^{(i)}\theta) - y^{(i)})\cdot X_1^{(i)} \\ \sum_i^m(\sigma(X_b^{(i)}\theta) - y^{(i)})\cdot X_2^{(i)} \\ \cdots\cdots\cdots\cdots\cdots \\ \sum_i^m(\sigma(X_b^{(i)}\theta) - y^{(i)})\cdot X_m^{(i)} \end{pmatrix}$$

,according to the vectorization process of linear regression, we can get,

$$\frac{1}{m}\begin{pmatrix} \sum_i^m(\sigma(X_b^{(i)}\theta) - y^{(i)}) \\ \sum_i^m(\sigma(X_b^{(i)}\theta) - y^{(i)})\cdot X_1^{(i)} \\ \sum_i^m(\sigma(X_b^{(i)}\theta) - y^{(i)})\cdot X_2^{(i)} \\ \cdots\cdots\cdots\cdots\cdots \\ \sum_i^m(\sigma(X_b^{(i)}\theta) - y^{(i)})\cdot X_m^{(i)} \end{pmatrix} = \frac{1}{m}\cdot X_b^T\cdot(\sigma(X_b\theta) - y)$$

,and then gradient descent.

4. System Structure Design

The algorithm model is run on Hadoop big data platform, and the real-time data is displayed on the front page through Kafka. The front end uses vue. js technology, and the back end uses Springboot technology [11]-[13]. UI design includes four parts: number of patients display module, virus parameter setting module, model prediction data trend display module, and virus diffusion trajectory diagram module, as shown in Fig. 4.

4.1. Hadoop Big Data Platform

Hadoop is a distributed system infrastructure developed by the Apache Foundation. Users can develop distributed applications without understanding the underlying details of distribution. Make full use of the power of clusters for high-speed computing and storage. Hadoop implements a Distributed File System (HDFS), one of its components. HDFS has high fault tolerance and is designed to be deployed on low-cost hardware [14]-[16]. Moreover, it provides high throughput to access the data of applications, which is suitable for applications with large data sets. HDFS relaxes THE POSIX requirements and allows streaming access to data in a file system. The core design of Hadoop framework is HDFS and MapReduce. HDFS provides storage for massive data, while MapReduce provides computing for massive data [17].

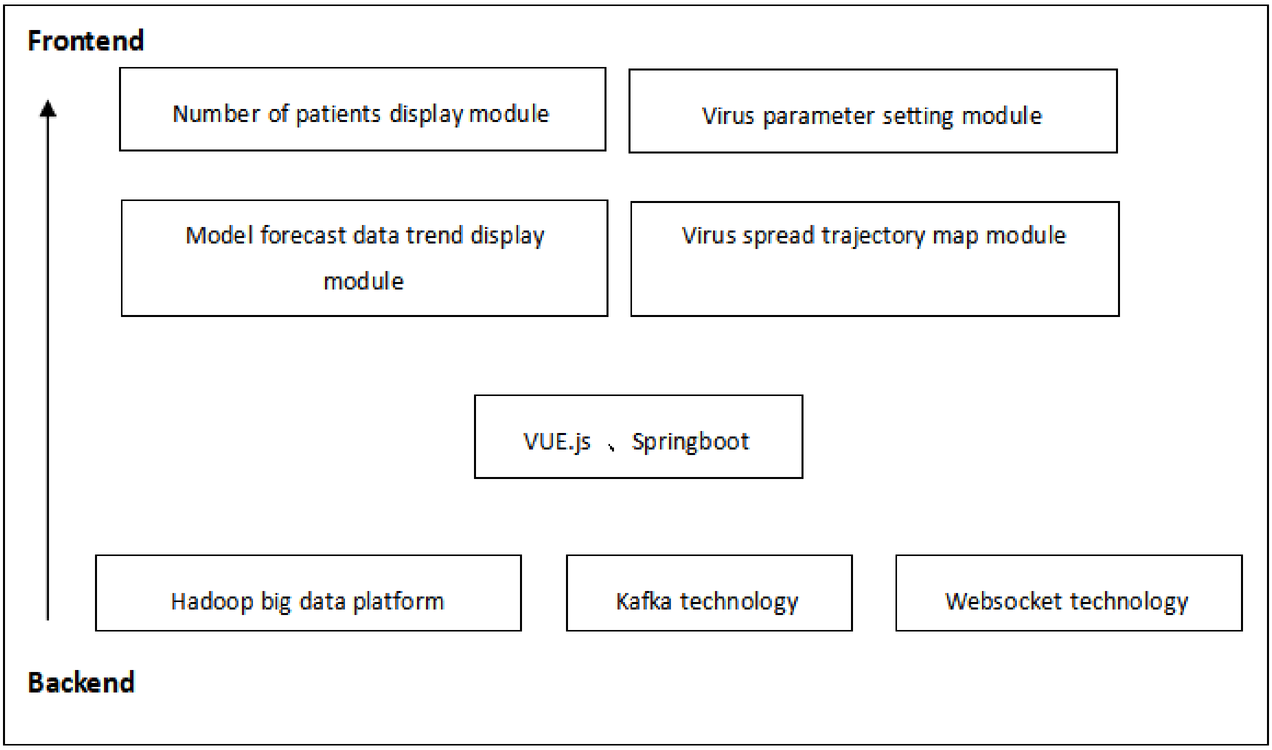

Frontend
Number of patients display module
Virus parameter setting module
Model forecast data trend display module
Virus spread trajectory map module
VUE.js 、Springboot
Hadoop big data platform
Kafka technology
Websocket technology
Backend

Fig. 5. System structure.

4.2. Kafka Technology

Kafka is an open source stream processing platform developed by the Apache Software Foundation and written in Scala and Java. Kafka is a high-throughput distributed publish-subscribe messaging system that processes all of the consumer's action-flow data in a website [18]-[19]. This action (web browsing, searching and other user actions) is a key factor in many social functions on the modern web. This data is usually addressed by processing logs and log aggregation due to throughput requirements. This is a viable solution for logging data and offline analysis systems like Hadoop that require real-time processing limitations. Kafka is designed to unify online and offline message processing through Hadoop's parallel loading mechanism, and to provide real-time messaging across clusters.

4.3. Number of Patients Display Module

The number of patients display module, mainly according to the S302 virus parameter Settings show the current number of patients.

4.4. Virus Parameter Setting Module

Virus parameter setting module, mainly to display virus parameters, including infectivity, latency, mortality and other parameters [20].

4.5. Model Prediction Data Trend Display Module

Model prediction data trend display module, mainly according to the virus parameter Settings to display the number of patients, new patients, cure and death data change trend [21].

4.6. Virus Diffusion Trajectory Diagram Module

Virus diffusion trajectory diagram module, mainly dynamic simulation of virus diffusion trajectory [22].

5. System Implementation

2019-NCOV epidemic development simulation system, as shown in Fig.5. Through the experiment, only the growth rate is calculated, the rest of the data are fitting, calculate the best r value was 0.18, end up with epidemic curve Fig. 6, the blue color squares represent outbreak confirmed infections, orange squares represent the recent infections predicted that represent the future of green squares infections predicted that the red curve represents infections fitting curve, The X axis represents the date, and the Y axis represents the number of confirmed cases [23], [24].

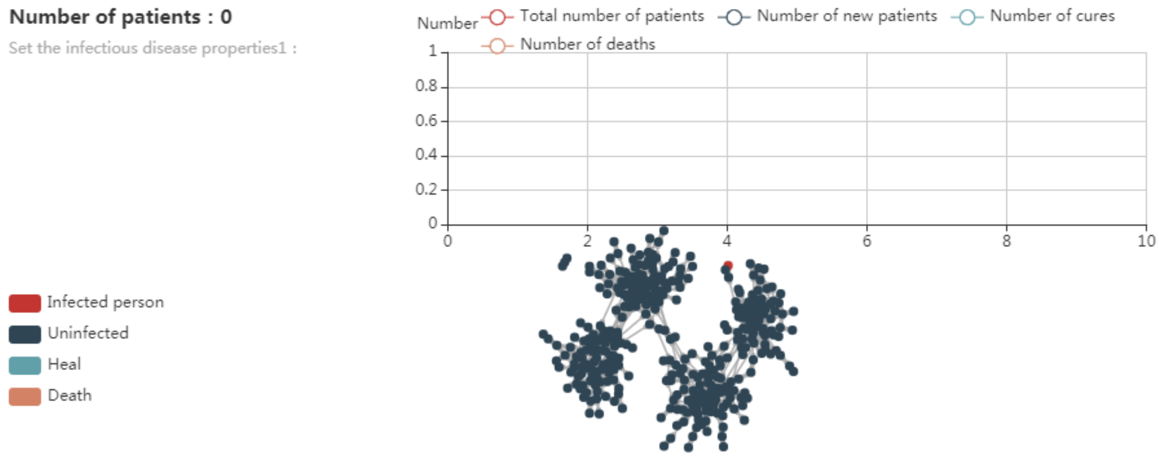


Fig. 5. 2019-NCOV epidemic development simulation system.

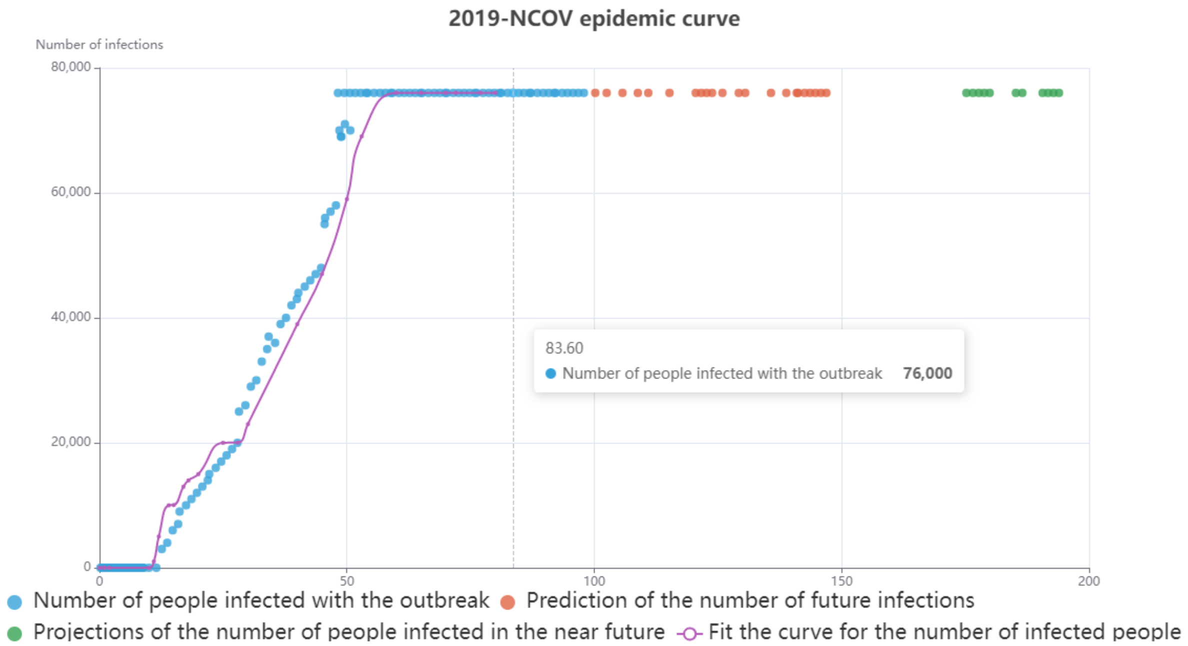


Fig. 6. Graph of the 2019-NCOV epidemic.

6. Conclusion

Epidemic development simulation platform, which simulates the epidemic development trend truly and effectively, provides high-quality data reference value for epidemic prevention and scientific research departments; 1) Use appropriate algorithms to make epidemic simulation more reliable and realistic; 2)

With intuitive, simple and easy to operate visual interface, to provide the best user experience for staff.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Dong-Xu Liu conducted the research; Feng Qin analyzed the data; Dong-Xu Liu completed the system implementation; Dong-Xu Liu wrote the paper; all authors had approved the final version.

Acknowledgment

The first author thanks that thorough preliminary investigation and analysis are essential for the subject being studied; The second author thanks that scientific rigor is something every scholar should keep in mind at all times; The third author believes that everything has its necessity. Of course, difficulties must accompany us. We should have the perseverance and determination to overcome obstacles.

References

[1] Nakimbugwe D., Ssepuuya G., Male D. *et al*. (2021). Status of the regulatory environment for utilization of insects as food and feed in Sub-Saharan Africa-a review. *Critical Reviews in Food Science and Nutrition*, *61(8).*

[2] Tewari, D. D. (2005). A brief historical analysis of water rights in South Africa. *Water International*, 2005, *30(4).*

[3] Mbzibain, A., & Tchoudjen, T. N. (2021). Ngo-state relations in the monitoring of illegal forest logging and wildlife trafficking in central Africa. *World Development*, *148*.

[4] Adekola, O., Krigsholm, P., & Riekkinen, K. (2021). Towards a holistic land law evaluation in sub-Saharan Africa: A novel framework with an application to Rwanda's organic land law 2005. *Land Use Policy*, *103*.

[5] Zhao J. H. (2018). A comment on "Factors influencing the outcome of primary immunization against rabies in young dogs. *Veterinary Microbiology*, 228.

[6] Konstantia, E. T. (2018). A reply to "A comment on "Factors influencing the outcome of primary immunization against rabies in young dogs." *Veterinary Microbiology*, *228*.

[7] Wilhelmus, G. P. M. (2018). Looijaard Letter to the editor: Comment on 'Timing of PROTein INtake and clinical outcomes of adult critically ill patients on prolonged mechanical VENTilation: The protinvent retrospective study'. *Clinical Nutrition*, *37(5)*.

[8] Morten, A. V. L., Anne, H. T., Christine, F. B. *et al*. (2020). Authors' reply to sert's comment on low-grade inflammation independently associates with cardiometabolic risk in children with overweight/obesity. *Nutrition, Metabolism and Cardiovascular Diseases*, *30*.

[9] Lai, C. H. (2022). Examining the paths of influence between individual motivators, information behaviors, and outcomes in disaster risk reduction. *International Journal of Disaster Risk Reduction*.

[10] Mwesiumo, D., Glavee, G. R., Olsen, K. M. *et al*. (2020). Improving public purchaser attitudes towards public procurement of innovations. *Technovation*.

[11] Silas, F. V. (2019). Digital textbooks are useful but not everyone wants them: The role of technostress. *Computers and Education*, 140.

[12] Zhu, W. W., Lu, S. S. (2021). Predicting public smog reduction behavior: Exploring the role of perceived risk and financial incentive policy. *Human and Ecological Risk Assessment: An International Journal*, *27(7)*.

[13] Nacarelli, V., & Gefen, D. (2021). trustworthiness and the adoption of business analytics. *Information Systems Management, 38(3)*.

[14] Wang, X. Y., Jia, D. Y., Gao, S. P. *et al*. (2020). Vaccination behavior by coupling the epidemic spreading with the human decision under the game theory. *Applied Mathematics and Computation*.

[15] Mo, B. C., Feng, K. R., Shen, Y. *et al*. (2021). Modeling epidemic spreading through public transit using time-varying encounter network. *Transportation Research Part C, 122.*

[16] Xi, B. B., Jiang, D. W., Li, S. H. *et al*. (2021). AutoVEM: An automated tool to real-time monitor epidemic trends and key mutations in SARS-CoV-2 evolution. *Computational and Structural Biotechnology Journal*.

[17] Guo, Z. Y., & Xiao, D. (2020). Analysis and prediction of the coronavirus disease epidemic in China based on an individual-based model. *Scientific Reports*, *10(1)*.

[18] Ge, Y., Zhang, W. B., Wang, J. H. *et al*. (2021). Effect of different resumption strategies to flatten the potential COVID-19 outbreaks amid society reopens: a modeling study in China. *BMC Public Health*, *21(1)*.

[19] Chen, Y. L., Li, Q., Karimian, H. *et al*. (2021). Spatio-temporal distribution characteristics and influencing factors of COVID-19 in China. *Scientific Reports*, *11(1)*.

[20] Oscar, V. M., & Marcel, M. A. N. L., Yorick, C. D. V. *et al*. (2018). Dynamics in copy numbers of five plasmids of a dairy lactococcus lactis strain under dairy-related conditions including near-zero growth rates. *Appl. Environ. Microbiol.*, *84(11)*.

[21] Liu, D. X., Feng, Q., Wang, Y. M., Li, X. R., Qian, Y., Zuo, D. D., & Zheng, Y. R. (2022). Design and implementation of graduate student integrated information management system — Degree management. *Journal of Software, 17(3)*, 121-129.

[22] Rinaldo, B. S. (2011). Strength in numbers? Not always! *Physica A: Statistical Mechanics and its Applications*, *390(10)*.

[23] Note on Cauchy's Numbers. *Annals of Mathematics*, *10(1/6)*.

[24] Elizabeth, R. A., & Julie, K. P. (2019). Strength in numbers: Mechanisms of viral co-infection. *Virus Research*, *265*.

D**ong-Xu Liu** is from the city of Ma'anshan in Anhui province, China. He was born in1991. His received his both undergraduate and postgraduate students graduated from the Computer Program of Anhui University of Technology with the following main research directions are computer applications and software engineering and big data applications.

He has received honorary title for several times from 2011-2017 for national motivational scholarship, top three students, excellent cadre, award for scientific and technological progress, first-degree scholarship, robot competition (first prize in province), provincial double superior;Between July 2017 and may 2021, he was employed at Siemens (China) Research Institute as a senior software development engineer while working as a project manager; Is primarily responsible for orbital traffic development and operation. Employed in Ma'anshan college in June 2021 as a computer professional teacher and title lecturer.

Mr. Liu, is a member of Anhui Provincial Computer Education Research Association.

Feng Qin is a professor and a master's tutor; he is currently the vice president of Ma on Shan College. A member of the 13th National People's Congress of Anhui province, Member of the Standing Committee of the 10th CPPCC of Maanshan City.

His main research interests are data mining, artificial intelligence, and computer networks.

In 2005, he received the "Baosteel Education Award" from the Baosteel Education Foundation.

In 2005, he was awarded the "Outstanding Instructor" by the Ministry of Education and the Ministry of Science and Technology, etc. In 2007,"Provincial Outstanding Teacher" by the Provincial Education Department.In 2013 "First Prize of Teaching Achievement" by the Provincial Education Department and he was awarded the "Provincial Master Teacher" by the Provincial Education Department.In 2015, he won the "Second Prize of Teaching Achievement" of the Provincial Department of Education.