PROCEEDINGS

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science

—— DCABES 2021 ——

10−12 December 2021
Nanning, China

PROCEEDINGS

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science

—— DCABES 2021 ——

10−12 December 2021
Nanning, China

Editor in Chief
Guo Yucheng

IEEE COMPUTER SOCIETY
CONFERENCE
PUBLISHING
SERVICES

BMS Part Number: CFP2120K-ART
ISBN-13: 978-1-6654-2889-7
ISSN: 2473-3636

Additional copies may be ordered from:

IEEE Computer Society
Customer Service Center
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1314
Tel: + 1 800 272 6657
Fax: + 1 714 821 4641
http://computer.org/cps
cps@computer.org

IEEE Service Center
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Tel: + 1 732 981 0060
Fax: + 1 732 981 9667
http://shop.ieee.org/store/
customer-service@ieee.org

IEEE Computer Society
Asia/Pacific Office
Watanabe Bldg., 1-4-2
Minami-Aoyama
Minato-ku, Tokyo 107-0062
JAPAN
Tel: + 81 3 3408 3118
Fax: + 81 3 3408 3553
tokyo.ofc@computer.org

Editorial production by Randall Bilof
Cover art production by Annie Jiu

IEEE COMPUTER SOCIETY
CONFERENCE
PUBLISHING
SERVICES

IEEE Computer Society
***Conference Publishing Services* (CPS)**
http://www.computer.org/cps

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

DCABES 2021

Table of Contents

Intelligent Transportation

Image and Vision Processing

Algorithm Design and Implementation

Network Security

Other Topics

DCABES 2021 Preface

The DCABES is a community working in the area of Distributed Computing and Applications in Business, Engineering, and Sciences, and is responsible for organizing meetings and symposia related to the field. DCABES intends to bring together researchers and developers in the academic field and industry from around the world to share their research experience and to explore research collaboration in the areas of distributed parallel processing and applications.

The 20th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES 2021) will be held on December 10~12, 2021 in Nanning, the metropolis of Guangxi Zhuang Autonomous Region. All papers accepted by DCABES 2021 Proceedings have been peer reviewed and carefully modified.

The DCABES series began as a summer short course held at Hong Kong Polytechnic University in 2000 with the support of the British Computer Society - Hong Kong Chapter. The two co-chairs of DCABES, Professors GUO Qingping and LAI Choi-Hong, extended the short course into a series of conferences that continues today and grows yearly.

We would like to thank all members of the Program Committee, the local organizing committee, and the external reviewers for selecting papers. We would also like to thank the WUT (Wuhan University of Technology, China), SLME (School of Logistics Management and Engineering, Nanning Normal University, China), the I^2C^3 (Institute of Intelligent Computing, Communication and Control, Wuhan) for their supports as local organizers of the conference.

Since the Covid-19 pandemic is not over until now, we take an online virtual conference to prevent spread of the disease. That is a new approach for academic exchange. I bless the DCABES2021 will be successful in advance.

Guo Qingping, Wuhan University of Technology, China
DCABES 2021 Cochair and I^2C^3 Chair

DCABES 2021 Conference Organization

Conference Chair

Guo Q. P., *Wuhan University of Technology, China*

Conference Cochairs

Douglas Craig C., *University of Wyoming, USA*

Lai C.-H., *University of Greenwich, United Kingdom*

Steering Committee

Craig C. Douglas*, University of Wyoming, USA*
Qing-Ping Guo*, Wuhan University of Technology, Wuhan, China*
Choi-Hong Lai*, University of Greenwich, UK*
Thomas Tsui*, Chinese University of Hong Kong, Hong Kong, China*
Wen-Bo Xu*, Jiangnan University, Wuxi, China*

Program Committee Cochairs

Craig C. Douglas, *Yale University, USA*

C.-H. Lai, *University of Greenwich, United Kingdom*

Q.P. Guo, *Wuhan University of Technology, Wuhan, China*

Yuhua Liu*, Hua Zhong Normal University, China*

Publication Chair

Yuchen Guo, *Wuhan University of Technology, Wuhan, China*

Scientific Committee (Paper Reviewer Bank)

Prof. Craig C. Douglas, *University of Wyoming, USA*
Prof. C.-H. Lai, *University of Greenwich, UK*
Prof. Q.P. Guo, *Wuhan University of Technology, China*
Prof. Frederic Magoules, *CentraleSupelec, Universite Paris-Saclay, France*
Prof. Albert Y. Zomaya, *The University of Sydney, Australia*
Prof. Maurício Vieira Kritz, *National Lab. for Scientific Computation, Petropolis-RJ, Brazil*
Prof. Peter Jimack, *University of Leeds, UK*
Prof. W.B. Xu, Jiangnan *University, Wuxi, China*
Prof. Xiao-Chuan Cai, *University of Colorado at Boulder, USA*
Prof. Jianwen CAO, *Institute of Software Chinese Academy of Sciences, Beijing, China*
Prof. Chi XueBing, *Chinese Academy of Sciences, China*
Prof. Souheil Khaddaj, *Kingston University, UK*
Prof. Yakup Paker, *Queen Mary University of London, London, UK*
Dr. Turgay Altilar, *Istanbul Technical University, Istanbul, Turkey*
Prof. David Keyes, *King Abdullah University of Science and Technology, USA*
Prof. Chen Wei, *Wuhan University of Technology, Wuhan, China*
Prof. Xiao Xinping, *Wuhan University of Technology, C*hina

Prof. Wenjing Li, *Guangxi Normal University, Nanning, Guangxi, China*
Prof. Shesheng Zhang, *Wuhan University of Technology, China*
Prof. Ping Lin, *University of Dundee, Dundee, UK*
Prof. Michael K. Ng, *The University of Hong Kong, HK*
Prof. Sun Jiachang, *Academy of Science, China*
Dr. Alfred Loo, *Lingnan University, HK*
Dr. Abal-Kassim Cheik Ahamed, *CentraleSupélec, Université Paris Saclay, France*
Prof. Peter Kacsuk, Xiao, *Hungarian Academy of Sciences, HU*
Prof. Stefan Vandewalle, *Katholieke Universiteit Leuven, Belgium*
Dr. Robert Lovas, *Hungarian Academy of Sciences, Hungary*
Dr. Faouzi Alaya Cheikh, *Gjovik University College, Norway*
Prof. NIKOS CHRISTAKIS, *University of Crete, Heraklion, Greece*
Prof. Haixin Lin, *Delft University of Technology, Netherlands*
Prof. Rassul Ayani, *Royal Institute of Technology, Sweden*
Prof. Zhihui Du, *Tsinghua University, China*
Prof. Meiqing Wang, *Fuzhou University, Fuzhou, China*
Prof. Yuhua Liu, *Central China Normal University-Wuhan, China*
Prof. Youwei Yuan, *Hangzhou Dianzi University, Hangzhou, China*
Prof. V. P. Kutepov, *Moscow Power Engineering Institute, Russia*
Prof. Yi Pan, *Georgia State University, USA*
Prof. Peter Sloot, *University of Amsterdam, Netherlands*
Prof. Xiaojun Tong, *Harbin Institute of Technology at Wei Hai, China*
Prof; Dan Liu, *Criminal Police University, China*
Dr. Haiwu He, *Hohai University, China*
Prof. Ziyue Tang, *Air Force Radar Institute, Wuhan, China*
Prof. Qifeng Yang, *Wuhan University of Technology, Wuhan, China*
Prof. Dongwoo Sheen, *Seoul National University, Seoul, Korea*
Prof. Rule Hjelsvold, *Gjovik University College, Norway*
Prof. Mohamed Kamel, *University of Waterloo, Canada*
Prof. Xinming Tan, *Wuhan University of Technology, China*
Prof. Shu Gao, *Wuhan University of Technology, China*
Dr. Professor Hongxing Liu, *Wuhan University of Technology, China*
Dr. Yucheng Guo, *Wuhan University of Technology, Ch*ina
Prof. Xiaojun Wu, *Jiangnan University, China*

Local Organization

Wenjin Li, Professor, *Nanning Normal University, China*
Qing-Ping Guo, Professor, *Wuhan University of Technology, China*
Dr. Yucheng Guo, *Wuhan University of Technology, China*
Dr. Chun Liu, *Wuhan University of Technology, China*

Secretariat

Dr. Yucheng Guo, *Wuhan University of Technology, China*
Dr. Chun Liu, *Wuhan University of Technology, China*

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Modification terms to the Black-Scholes model based on the functional volatility model

Xiaozheng Lin*, Wei Chen, Meiqing Wang
College of Mathematics and Computer Science
Fuzhou University
Fuzhou, China
Email: lynn.xiaozheng@gmail.com;965340978@qq.com;
mqwang@fzu.edu.cn

Choi-Hong Lai
Department of Mathematics
University of Greenwich
London,SE10 9LS,UK
Email: C.H.Lai@greenwich.ac.uk

***Abstract*—The Black-Scholes model (B-S model) is a common used option pricing model, which assumes that the volatility and the interest rate are constants. However, the empirical analysis shows that the option prices obtained from the B-S model are quite different from the market prices. The traditional improvement to the B-S model is to replace the volatility constants or interest rate constants with variables. This paper proposes the concept of modification term of Black-Scholes model, which is different from the traditional methods.The errors of the B-S model may be calculated through the market data, and is fitted as a function of a functional implied volatility and the option gamma. The fitted function can be considered as the modification term of the Black-Scholes model. The functional implied volatility can be constructed by using the Gaussian semi-parametric implied volatility model proposed by Ref. [4]. Experimental results show that the modified model has a better predictive effect on option pricing.**

***Index Terms*—right-term values;modification term;functional implied volatility;modified Black-Scholes model**

I. Introduction

In the early 1970s, Fischer Black and Myron Scholes proposed the first complete option pricing model, deriving the famous Black-Scholes option pricing model (also known as the Black-Scholes model) under a series of strict assumptions [1].

The Black-Scholes model requires the assumption that the underlying asset volatilities are constant [1]. It is practically impossible to execute in real markets. Therefore, option pricing results using constant volatilities in the Black-Scholes partial differential equation (PDE) are inevitably to be different from pricing results using implied volatilities [2]. The objective of this paper is to address this problem through the use of functional volatility model instead of the constant historical volatility to generate the modification term of the Black-Scholes model.

The paper is organised as follows. Section II discusses the modification term to the Black-Scholes model when an suitable implied volatility model is used. Section III provides a functional implied volatility model and introduces the modification term based on the functional implied volatility model we established. Section IV gives the corresponding numerical experiments and analysis of the method proposed in the previous. Finally, conclusions are drawn.

II. Modification Term of The Black-Scholes Model

A. Black-Scholes Model and Its Modification Term

Under a series of strict assumptions, the relationship between option price, underlying asset price and time can be described by the following famous Black-Scholes partial differential equation(PDE) [1]:

$$\frac{\partial V}{\partial t} + rS\frac{\partial V}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} = rV, \tag{1}$$

subject to suitable terminal and boundary conditions provides deterministic price of many options and derivatives. Here S denote the underlying asset price, V is the option price, r is the risk-free rate and σ means the underlying asset volatility. The volatility of the underlying asset calculated inversely from the price of a single option becomes the implied volatility [1]. The Black-Scholes model assumes that the volatility of the underlying asset σ is constant.

Perform simple item shift processing on model (1):

$$\frac{\partial V}{\partial t} + rS\frac{\partial V}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} - rV = 0. \tag{2}$$

It can be found from equation (2) that if the Black-Scholes model has no errors at all, then substituting the actual option price into the left of equation (2) for derivative calculation, then the right-term should be a constant 0 [3].

However, a large number of numerical experimental studies have found that the right term is non-zero, which shows that the B-S model under strict assumptions does not reflect the actual market conditions well and the non-zero right term can be considered as the errors of the B-S model for the actual market [3].

As shown in Fig.1, taking the 2019 SSE 50ETF purchase of December 2950 option as an example, substituting the actual market price data of this option into equation (2) to obtain the logarithm value of the right-term.

In this paper, the right-term values are used as the modification term f of the Black-Scholes model:

$$\frac{\partial V}{\partial t} + rS\frac{\partial V}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} - rV = f. \tag{3}$$

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00007

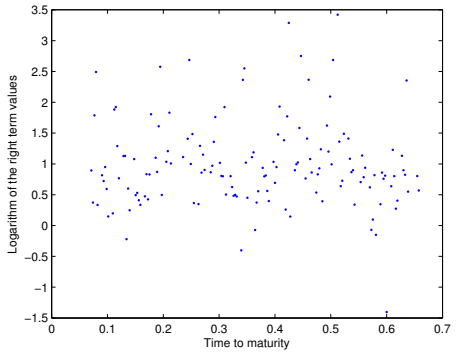


Fig. 1. Logarithm of the right term

Therefore, if we can model the right-term f calculated from the actual transaction data, it is equivalent to finding the modification term function of the B-S model, and the original Black-Scholes model can be modified according to the modification term.

B. Modification Term Based On Implied Volatility Model

The assumption that the volatility of the underlying asset is a constant is one of the reasons for the error of the traditional Black-Scholes model. Therefore, in order to model the actual right-term f and amend the Black-Scholes model, the most direct solution is to use an implied volatility model that can characterize the time-varying characteristics of underlying asset volatilities to derive the modification term of the Black-Scholes option pricing formula.

Since the implied volatility of the underlying asset price is calculated by substituting the actual option price into the Black-Scholes model, then equation(4) can be obtained when a suitable model (4) is considered to be a good fit for the actual implied volatility. Replace σ in equation(3) with the suitable implied volatility model $\widehat{\sigma}$:

$$\frac{\partial V}{\partial t}+rS\frac{\partial V}{\partial S}+\frac{1}{2}\widehat{\sigma}^2S^2\frac{\partial^2V}{\partial S^2}-rV=0. \tag{4}$$

The difference of equation (3) minus equation (4) is used as the modification term f of Black-Scholes model,

$$f=\frac{1}{2}\left(\sigma^2-\widehat{\sigma}^2\right)S^2\frac{\partial^2V}{\partial S^2}. \tag{5}$$

Therefore, the problem of B-S model modification from the perspective of right-term modelling is transformed into a problem of finding a better implied volatility model $\widehat{\sigma}$ to generate modification term equation(5).

III. The Modification Term Based on The Functional Implied Volatility Model

A. The Gaussian semi-parametric implied volatility model

Contrary to the assumption that the implied volatility is constant in the Black-Scholes model, many scholars have found that implied volatility has the structural characteristics of smile, smirk, and skew in relation to execution price.On the other hand,many studies also show the relationship between implied volatility and the maturity date of an option.

In order to better express the degree of curvature and translation position of the implied volatility delivery structure, Ref. [4] used the inverse Gaussian function and parametric exponential variables to improve the implied volatility surface model proposed by Cassese & Guidolin [5],and obtained the Gaussian semi-parametric implied volatility model:

$$\begin{aligned}\widehat{\sigma}_\tau(M)=&x_{1,\tau}+x_{2,\tau}\left(M-x_{3,\tau}\right)\\&+x_{4,\tau}[\max(\frac{1}{\theta\sqrt{2\pi}}e^{-\frac{(M-x_{3,\tau})^2}{2\theta^2}})-\frac{1}{\theta\sqrt{2\pi}}e^{-\frac{(M-x_{3,\tau})^2}{2\theta^2}}]^{\alpha_\tau}\\&+\varepsilon_\tau,\end{aligned} \tag{6}$$

here M denote the moneyness:

$$M=\ln\left(\frac{S}{K}\right), \tag{7}$$

$x_{1,\tau}$ and $x_{3,\tau}$ are parameters that determine the translational position of the fitted curve,$x_{3,\tau}$ is the mathematical expectation of M.$x_{2,\tau}$,$x_{4,\tau}$ and ε_τ are parameters that determine the bending amplitude of the fitted curve.θ is the variance of M. They are all fitted and calculated from market data [4].

B. Functional Implied Volatility Model

Model (6) can well simulate the implied volatility when the remaining period is τ in a certain trading day t [4]. In other words, model (6) fixes the time to maturity τ, and only considers the influence of the moneyness M within each τ. Functional implied volatility model proposed in this paper provides a model which can consider the influence of both moneyness M and the time to maturity τ:

$$\begin{aligned}\widehat{\sigma}(\tau,M)=&f_1\left(\tau\right)+f_2\left(\tau\right)\left(M-\mu_M\right)+f_3\left(\tau\right)[\max(\frac{1}{\theta\sqrt{2\pi}}\\&e^{-\frac{(M-\mu_M)^2}{2\theta^2}})-\frac{1}{\theta\sqrt{2\pi}}e^{-\frac{(M-\mu_M)^2}{2\theta^2}}]^{\alpha_\tau}+\varepsilon_\tau,\end{aligned} \tag{8}$$

μ_M represents $x_{3,\tau}$ in equation (6), which is the mean value of moneyness M.

In order to construct the randomness of the implied volatility surface, a fourier function is used to replace the constant $x_{1,\tau}$ in the case of a fixed remaining term τ:

$$f_1\left(\tau\right)=a_0+a_1\cos\left(w*\tau\right)+b_1\sin\left(w*\tau\right). \tag{9}$$

Consider the randomness and asymmetry in the τ direction:

$$f_2\left(\tau\right)=a_2\sin\left(b_2\tau+c_2\right), \tag{10}$$

$$f_3\left(\tau\right)=a_3\exp\left(b\tau\right), \tag{11}$$

here $f_2\left(\tau\right)$ denote the function sum of sine, and $f_3\left(\tau\right)$ is a exponential function.Equations $(8)-(11)$ constitute the functional implied volatility model proposed in this paper.

C. The Modification Term Based on The Functional Implied Volatility Model

The functional implied volatility model $(8)-(11)$ is used as a suitable implied volatility model of equation(4). According to equation(5),the modification term of the Black-Scholes model should be:

$$f = \frac{1}{2}\left(\sigma^2 - \widehat{\sigma}^2\left(\tau, M\right)\right) S^2 \frac{\partial^2 V}{\partial S^2}. \tag{12}$$

Therefore, the Black-Scholes model with modification term function (12) is:

$$\begin{aligned} &\frac{\partial V}{\partial t} + rS\frac{\partial V}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} - rV \\ &\qquad = \frac{1}{2}\left(\sigma^2 - \widehat{\sigma}^2\left(\tau, M\right)\right) S^2 \frac{\partial^2 V}{\partial S^2}, \end{aligned} \tag{13}$$

σ represents the constant volatility of the underlying asset in the original Black-Scholes model (1) and $\widehat{\sigma}\left(\tau, M\right)$ are the implied volatilities calculated from the functional implied volatility model(8)-(11) under a certain time to maturity τ and a moneyness M.

Therefore, equations(8)-(13) constitute the modified Black-Scholes model.

IV. Numerical Experiments and Analysis

A. Data Preprocessing

This paper examines 9 European-style call options on the SSE 50ETF in December 2019. Under every different remaining terms $t_{ij}, i = 1, 2, ..., 9, j = 1, 2, ..., n_i (t_{i1} = 0, t_{iN} = T)$, these 9 options had the same maturity date T, risk-free interest rate r and the same underlying asset prices $S(t_{ij})$ with different strike prices K_i.

The numerical experimental part of this section will show and analyze the fitting results of the functional implied volatility model, the construction accuracy of the modification term and the forecast accuracy of option pricing on the SSE 50ETF in December 2020.

All data are sourced from the Cathay Financial Database.

B. The Fitting Results of The Functional Implied Volatility Model

The following figure 2 shows the functional implied volatility surface (8)'s fitting results of 9 options on the SSE 50ETF in December 2020.

It can be seen from Fig.2 that the functional implied volatility surface model constructed in this paper can fit the actual implied volatilities well.

C. The Construction Accuracy of The Modification Term

In the numerical experiment of this paper, the modification term values calculated by equation (12) for 9 options were compared with the actual modification term values f_{true} which were calculated by equation (3) using the actual market data.

From the comparison results, we can observe the accuracy of the modification term equation (12) of the Black-Scholes model proposed in this paper. Consider using the Euclidean

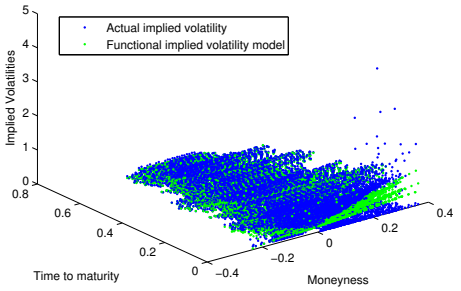


Fig. 2. Fitting results of functional implied volatility models for 9 European call options

distance as a measure of the modification term values obtained by equation (12) and the actual right-term values:

$$d = \sqrt{\sum_{i=1}^{N} \left(f\left(t_i\right) - f_{\text{true}}\left(t_i\right)\right)^2}, \tag{14}$$

this paper selects the minimum, maximum and average Euclidean distances of the 9 options used in the experiment to illustrate the accuracy of the modification term equation (12).

TABLE I
THE CONSTRUCTION ACCURACY OF THE MODIFICATION TERM

Euclidean distance	minimum	maximum	average
Values	0.1350	2.3980	1.2696

It can be seen from Table 1 that the fitting degree between the modification term values f obtained by equation (12) and the actual modification term values f_{true} is relatively high.

D. The Performance of The Modificated Black-Scholes Model on the Accuracy of Option Prices

The experiment used the modified Black-Scholes model (8)-(13) to predict the 2020 option prices corresponding to the 9 options used in the experiment, and compared the predicted results with the actual option prices in 2020.

Fig.3 shows the comparison between the results of the modified Black-Scholes model (8)-(13) and the actual option prices in 2020(4 out of 9 options).

It can be seen from Fig.3 that the modified Black-Scholes model results are closer to the actual values than the traditional Black-Scholes model results.

E. Model Analysis and Evaluation

The main purpose of this paper is to establish the modification term of the Black-Scholes model from the right-end error calculated by substituting the actual market option price into the Black-Scholes equation(3), and at the same time try to use the modified model to predict option prices.

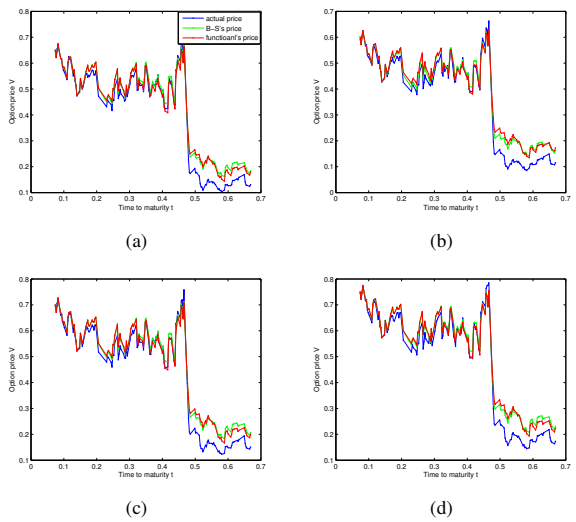


Fig. 3. Results of the modified Black-Scholes model on the accuracy of option prices

From the numerical experiment results, it can be found that the simulation results of the modification term model (8)-(12) proposed in this paper can fit the right-end errors of the actual model well,but lacking in predicting option prices.This reminds us that we can study the modification term and prediction accuracy of the Black-Scholes model from the aspects of the computational stability and well-posedness of the differential equations (8)-(13).

But even so, this paper provides a novel idea for the modification of the Black-Scholes model, which is innovative.And under this idea, a modification term with a high degree of fit is constructed, so that the follow-up work can be carried out based on this modification term or the modification idea.

V. Conclusion

This paper proposes the concept of the modification term of the Black-Scholes model, and uses actual market transaction data to calculate the error values at the right-term of the Black-Scholes equation (3), and then establishes the modification term model (8)-(12) of the Black-Scholes model.In this paper, numerical experiments are designed to start with the accuracy of the construction of the modification term model (8)-(12) and the accuracy of using the modified equation (8)-(13) to predict option prices.

The empirical analysis shows that the modification term model (8)-(12) can fit the actual right-end errors of the Black-Scholes equation well. At the same time, the modified Black-Scholes model can also predict option prices more accurately.

References

[1] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.

[2] S Chandra Sekhara Rao et al. Numerical solution of generalized black–scholes model. *Applied Mathematics and Computation*, 321:401–421, 2018.

[3] Choi-Hong Lai. Modification terms to the black–scholes model in a realistic hedging strategy with discrete temporal steps. *International Journal of Computer Mathematics*, 96(11):2201–2208, 2019.

[4] Xiaoyan Wu, Ying Zhuang, Fei Chen, and Meiqing Wang. A gaussian semi-parametric implied volatility model. *Journal of Algorithms & Computational Technology*, 11(3):246–260, 2017.

[5] Gianluca Cassese and Massimo Guidolin. Modelling the implied volatility surface: Does market efficiency matter?: An application to mib30 index options. *International review of financial analysis*, 15(2):145–178, 2006.

[6] Seda Gulen, Catalin Popescu, and Murat Sari. A new approach for the black–scholes model with linear and nonlinear volatilities. *Mathematics*, 7(8):760, 2019.

[7] James O Ramsay. Functional data analysis. *Encyclopedia of Statistical Sciences*, 4, 2004.

[8] James O Ramsay and Bernard W Silverman. *Applied functional data analysis: methods and case studies*, volume 77. Springer, 2002.

[9] Germán Aneiros, Ricardo Cao, Ricardo Fraiman, Christian Genest, and Philippe Vieu. Recent advances in functional data analysis and high-dimensional statistics. *Journal of Multivariate Analysis*, 170:3–9, 2019.

[10] Samuel M Nuugulu, Frednard Gideon, and Kailash C Patidar. A robust numerical scheme for a time-fractional black-scholes partial differential equation describing stock exchange dynamics. *Chaos, Solitons & Fractals*, 145:110753, 2021.

[11] Zeyad Al-Zhour, Mahdiar Barfeie, Fazlollah Soleymani, and Emran Tohidi. A computational method to price with transaction costs under the nonlinear black–scholes model. *Chaos, Solitons & Fractals*, 127:291–301, 2019.

Empirical Research on the Impact of Margin Selling on the Volatility of China's Stock Market

Gao Xi
Guilin University of Electronic Technology
GuiLin, China
e-mail: 919784364@qq.com

Song Ruimin
Guilin University of Electronic Technology
GuiLin, China
919784364@qq.com

***Abstract*—In March 2010, China's securities market began the trial of margin trading and ended the unilateral trading system since the operation of the stock market in China. It is an important milestone in the development of China's securities market. From 2010 to now, the scale of margin trading continues to expand, the quality of the underlying stocks is getting higher and higher, the regulation of the market is becoming more and more perfect, but how does it affect the volatility of the stock market in China? So far there is no uniform conclusion.**

This paper deals with the daily trading data of the Shanghai and Shenzhen 300 Index as an indicator of the volatility of the stock market, and uses the total financing balance and margin of the Shanghai and Shenzhen Stock Exchange as the margin trading index. This paper makes an empirical analysis of the impact of margin trading on the volatility of China's stock market. Samples range from March 31, 2010 to March 31, 2016, a total of 1643 trading days.

Keywords-component;:VAR Model Volatility Stock Market Margin trading

I. INTRODUCTION

Since margin trading has a very mature trading system in foreign countries, the research based on the stock market has a certain system, while the development of margin selling trading in China is still immature. Therefore, the research based on China's stock market is relatively few.At present, scholars at domestic and foreign countries have not had a relatively consistent view on "the impact of margin trading on the stock market". The research conclusions are roughly divided into: 1. Margin trading will aggravate the stock market volatility; 2. Margin trading will inhibit the stock market volatility; 3. Margin selling has no impact on the stock market.The literature for these three different results is presented separately below:

II. RELATED THEORY

2.1 Relevant concepts of margin selling

2.1.1 The definition of margin selling, leverage

Margin selling trading is also credit trading, investors only pay a certain margin to enlarge their trading chips, but also expand the transaction of profits and losses several times.Investors through the judgment of the market, short buy short trading, although its own funds or securities is not much, but using the leverage of margin trading, investors only pay a small part of margin can obtain several times the number of investment funds or securities

2.1.2 The features of margin selling

(1) The flow of funds

As an organic part of the financial market, the money market and the capital market must maintain a smooth capital flow between the two markets to ensure the smooth operation of the financial market. If the circulation of funds between the two markets is blocked, it will inevitably have an impact on the function of the whole financial market.The margin trading mechanism determines its contribution to capital circulation.

(2) Leverage

The action mechanism of margin trading on stock market volatility.Margin trading is mainly through its leveragemechanism and short selling mechanism.Leverage mechanism means that investors get over a small percentage of trading chips to pay a margin, thus increasing the supply of the securities market.Short selling mechanism happens when investors borrow shares at high levels to sell when the stock price is above their intrinsic value.

(3) The forward mechanism of action

When the market is in a fully effective market, when the stock price is seriously overvalued, rational investors will use short selling, margin selling stock, increase the supply of stock, hinder the stock rise, buy securities to repay, and under the influence of sheep effect, other investors will sell stock at high.

(4) Negative action mechanism

In the case that the market is weak and effective, a large number of irrational investors do not grasp the internal information of listed companies, easy to produce overreaction, that is, investors show excessive optimism about good news, show extreme pessimism about negative news, resulting in a certain cognitive deviation.(1) Total ban phase (1990-2005)

III. MPIRICAL STUDY ON THE EFFECT OF MARGIN ON STOCK MARKET VOLATILITY

3.1 Selection of samples and data sources

The starting date is the starting date, and the data interval from March 3 1-30-10 December 2016, a total of 1643 groups were selected as the empirical study samples.The data selected in this article are all obtained from the Resset financial database. Eviews software was used to construct a VAR vector autoregression model to test the experimental hypothesis.This paper sets the variables involved in the empirical study and selects the study indicators as follows:

The volatility of the stock market.The daily volatility of the CSI(China Securities Index) 300 index, namely the ratio between the daily highest and lowest price and the daily average daily price, is recorded as the calculation formula adopted by VOL(Volume), as follows:

$$\mathrm{VOL} = 2(\mathrm{P_t^H} - P_t^L)/(\mathrm{P_t^H} + P_t^L) \quad (1)$$

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00008

P^H_t represents the highest price for the CSI 300 t rading on t-day, and P^L_t says the CSI 300 t rades at the lowest price on t daily

Financing transaction balance.The total financing balance of Shanghai and Shenzhen was selected. Due to the large transaction amount, in order to minimize the error of empirical analysis, the total financing balance was processed logarithmically treated here to eliminate heterovariance and recorded as lnMP(Financing transaction balance).

Margin trading balance.The total margin margin for each trading day between Shanghai and Shenzhen were selected, and the total margin was also log-treated to eliminate heterovariance and be recorded as lnSS(Margin trading balance).

3.2 Stability test

The stationarity of time series should first be considered first in the process of selecting variables, but in practice, time series variables are most likely to be unstable. If the regression of the data without stationarity test is conducted directly, it is easy to get false conclusions (pseudo-regression relationship).Therefore, when conducting the time-series study, the unit root test of the studied time-series variables should be first carried out to determine whether it is stable.This paper performs ADF(Augmented Dickey-Fuller test) tests for the data using Eviews9.0.

Variable name	ADF statistics	1% Critical value	5% Critical value	Pvalue	conclusion
VOL	-7.959608	-3.963656	-3.412555	0.0000	Refusing the original hypothesis, Smooth
LNMP	-5.236791	-3.963659	-3.412557	0.0001	Refusing the original hypothesis, Smooth
LNSS	-4.904297	-3.963662	-3.412558	0.0003	Refusing the original hypothesis, Smooth

Table-1

From Table -1 available at the 1% significance level, the absolute value of the ADF statistic of all three variables is significantly greater than the absolute value of the cutoff, so the null hypothesis can be rejected, all stationary number columns.

3.3 The VAR model was established and the optimal lag order was determined Estimation of the VAR modelIn

1980, Sims proposed a vector autoregressive model, simply the VAR model.This model takes the form of multiple equations, in each equation of the model, the endogenous variable regressed the lag values of the model, which estimates the dynamic relationship of all endogenous variables.In this paper, a binary VAR model is established to regression analyze the relationship between stock market volatility and financing trading and margin trading respectively. The expressions of the VAR model are as follows:

$$VOL_t = c_1 + \beta_1 VOL_{t-1} + ... + \beta_k VOL_{t-k} + \alpha_1 LnMP_{t-1} + ... + \alpha_i LnMP_{t-i} + \mu_t \quad (2)$$

$$LnMP_t = c_2 + \lambda_1 LnMP_{t-1} + ... + \lambda_i LnMP_{t-i} + \theta_1 VOL_{t-1} + ... + \theta_j VOL_{t-j} + \mu_t \quad (3)$$

$$VOL_t = c_3 + \alpha_1 VOL_{t-1} + ... + \alpha_i VOL_{t-i} + \beta_1 \mathrm{Ln}SS_{t-1} + ... + \beta_j LnSS_{t-j} + \mu_t \quad (4)$$

$$LnSS_t = c_4 + \lambda_1 LnSS_{t-1} + ... + \lambda_i LnSS_{t-i} + \theta_1 VOL_{t-1} + ... + \theta_j VOL_{t-j} + \mu_t \quad (5)$$

3.3.1Determine the lag order of the VAR model

Before the final determination of the VAR model, the size of the lag order of the model will affect the accuracy of the model estimation results, if the selected lag order is small will make the autocorrelation of the model error term is relatively large, but if the lag order is too large will also reduce the degree of freedom of the model, so the lag order of the model should not be too large or too small.

In this paper, the Lag Length Criteria rule in the Eviews software, namely the LLC rule, is chosen to test the optimal lag period number of the VAR model.The optimal number of lag periods shown is 7, and the two metrics of FPE, AIC is 8, finally identifying the VAR model as 7 according to LLC's test rules.

3.3.2 The stability of the VAR model was tested

After the optimal lag order of the model is determined, the stability of the VAR model should be tested, generally taking the AR root test, whose principle is that the VAR model is considered stable if all the feature roots of the tested VAR model are within the unit circle, i. e., only if the absolute value of the feature root is less than 1.

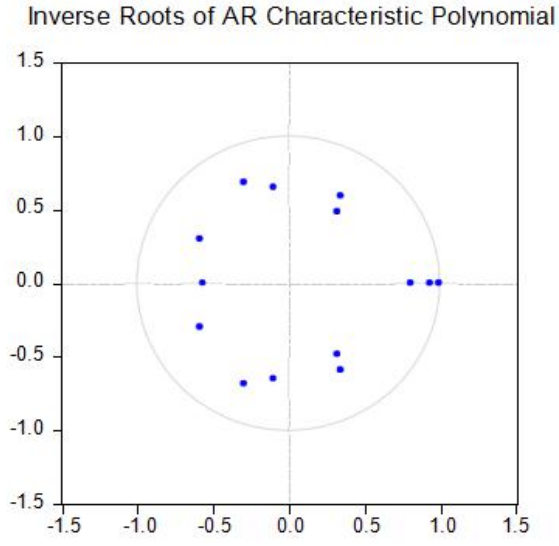


Figure-1

3.3.3 Establishment of the VAR model and the outcome analysis

From the regression coefficient and t values in the constructed VAR model, t he volatility of the CSI 300 index is significantly related with its own lag period and shows positive correlation; the volatility of the CSI 300 index is subject to the logarithm of the total financing balance.The impact of a significant negative correlation between both 2 and 6 lag, indicating that the change in financing balance suppressed the wave of the stock market.

3.4Granger.causal.test

Table-2

Dependen variable	Exclude	Chi-sq	df	Prob.
VOL	LNSS	6.15809	7	0.5263
LNSS	VOL	9.792104	7	0.2007
VOL	LNMP	13.17693	7	0.0198
LNMP	VOL	16.23778	7	0.0230

From the test results in the table, At the 5% significance level, Testing accepted "VOL is not Granger reason for LNSS" and "LNSS is not Granger reason for VOL", Also testing rejection of "VOL is not Granger reason for LNMP" and "LNMP is not Granger reason for VOL", That is, margin trading is not the Granger cause of stock market volatility, But financing deals are the Granger reason for stock-market volatility, And the volatility of the stock market is also the reason for margin trading.

3.5 Variance decomposition

Table-3

periods	VOL	LNMP	VOL	LNSS
1	100.0000	0.000000	100.0000	0.000000
2	99.99373	0.006270	99.89984	0.100157
3	99.69079	0.309210	99.62184	0.378161
4	99.68139	0.318606	99.61263	0.387372
5	99.66958	0.330416	99.62639	0.373614
6	99.66225	0.337750	99.58856	0.411440
7	99.67115	0.328849	99.59855	0.401452
8	99.68367	0.316328	99.60033	0.399668
9	99.69157	0.308426	99.59638	0.403625
10	99.69738	0.302619	99.59170	0.408299

Can be seen from the table, our stock market volatility is most affected by their own, its proportion of their interpretation accounted for more than 99%, and basically stable in the long term, and margin trading to stock market volatility is less than 1%, shows that margin trading on stock market volatility is not big, in the first ten lag period, the impact is not more than 1%, and margin trading than financing contribution to stock market volatility, this result is basically consistent with the results of the pulse response.

3.6 pulse response analysis

The pulse response is used to describe the response of an endogenous variable to the error impact, where he describing the impact on the current and future values of the endogenous variable after an impact of one unit on the random error term.The effect of the pulse is temporary and slowly approaches to zero over time.

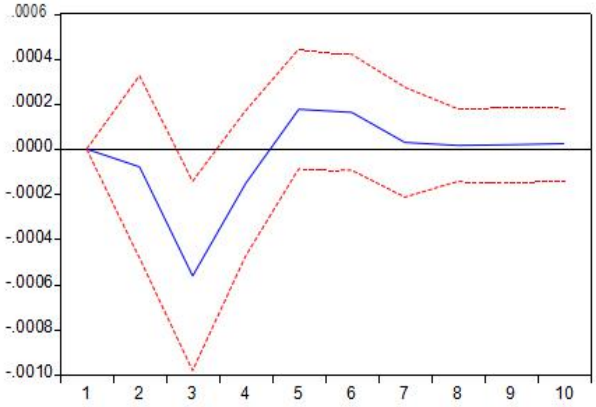


Figure-2 pulse response

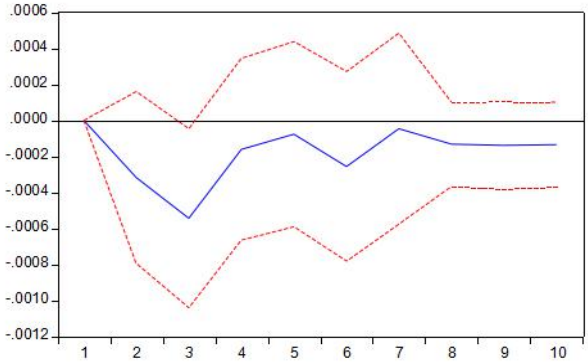


Figure-3 pulse response

After the positive impact of a unit of margin trading, the stock market showed a negative reflection in the first time, and gradually increased, gradually in the third period to the peak gradually reduced and fluctuated back and forth, and the final fluctuation gradually decreased and tends to stabilize.The above results show that margin trading will bring a negative impact on the stock market volatility of the stock market, that is, margin trading will inhibit the stock market volatility, but the effect will gradually decrease over time.

Ⅳ Conclusion and Policy Suggestions

Conclusions

This paper selects the margin trading data of CSI 300 index and CSI, selects appropriate variables, and constructs the VAR model to study the impact of margin trading on the volatility of the stock market.From the pulse response and variance decomposition, margin trading has a certain influence on the stock market volatility, in the whole research range, margin trading has suppressed the stock market volatility, namely the introduction of margin trading will reduce the volatility of the stock market, and margin trading on the stock market volatility is slightly greater than the impact of financing trading on the stock market volatility.

Suggestions

I. Further increase the securities and types of securities underlying margin selling.China's margin securities expanded four times to varying degrees from 2010 to 2016, but since the sharp fluctuations of the stock market in

2005, the CSRC has strengthened the regulation of securities.

II. Make flexible adjustments to the margin system for margin trading and securities. The development of margin trading and securities is mainly based on the margin system, and the margin ratio determines the level of leverage for short buying and selling. At the same time, the size of the margin ratio will also affect the investment choices of investors. If the regulator can adjust the margin ratio of margin trading and securities according to the current overall market operation, the market stability function of the two financial services will be greater.

REFERENCES

[1]Li Z, Lin B, Zhang T, et al. Does short selling improve stock price efficiency and liquidity? Evidence from a natural experiment in China[J]. The European Journal of Finance, 2016: 1-23.

[2]James[R], 1997.J Angel. Short selling on the NYSE[R],1997.

[3]Chang EC, Bai Y,Wang J.Asset Prices under Short-Sales Constraints[J].Working Paper,2006(11):1-43.

[4]Yang Deyong, Wu Qiong.Empirical Analysis of the Impact of Margin on Shanghai Securities Market — — Based on liquidity and volatility perspective [J].Journal of the Central University of Finance and Economics, 2011 (05): 28-34.

[5]Liao Shiguang, Yang Chaojun.Short selling trading mechanisms, volatility and liquidity — — An empirical study based on the Hong Kong stock market [J].Managing World, 2005 (12): 6-13 + 171.

[6]Xiao Wenyan, Wang Zihan.Influence of Margin Mechanism on Volatility of China Stock Market — — Based on Empirical Analysis of SSE 50 [J].China Securities and Futures, 2012 (05): 33-35.

E-commerce Service Equilibrium Prediction Simulation Based on Ontology

WU ZHONG
Physical Education School
Wuhan Business University
Wuhan, China
7849800@qq.com

Abstract— **Accurately predicting the needs of users in the process of e-commerce services can expand the scope of e-commerce services and improve the quality of e-commerce services. When forecasting demand, it is necessary to establish different forecasting indicators for different factors that affect e-commerce services. The traditional dual-neighbor selection strategy can only select recommended behaviors based on a single indicator, which reduces the demand for e-commerce services. Accuracy. This paper proposes an ontology-based method for predicting user demand balance in the process of e-commerce services. Firstly, construct an e-commerce service balanced recommendation user ontology, calculate the semantic similarity between the ontology, and obtain a set of similar users. On this basis, predict the neighbors of the target user as the prediction group, and calculate the trust subgroups with higher prediction probability in the group. Through the dynamic measurement of uncertain neighbors for balanced prediction, effectively completed the balanced prediction of user demand in e-commerce services. Simulation experiments prove that the proposed method can effectively improve the accuracy of user demand prediction in e-commerce services, and it is valuable.**

Keywords- E-commerce service, Ontology, Equilibrium Prediction, Simulation

I. INTRODUCTION

With the continuous development of mobile Internet technology, the amount of information on the Internet is increasing day by day, and the recommendation of business objects under the electricity supplier mode has attracted people's special attention [1]. As a new information service mode, demand forecast can effectively alleviate the problem of information overload [2]. However, in most e-commerce services, user demand forecasting system can not accurately calculate the similarity between users' ontology, accordingly reducing user demand prediction accuracy. Under such condition, improving the quality of e-commerce service demand forecast has been paid much attention by the relevant people, and some progress has been made.

However, there are many problems in existing research, such as large deviation [3], difficult to resist the interference of external factors [4], and can not accurately calculate the similarity between users [5].

In view of the problems arising above, it proposes a balanced recommendation method based on ontology for e-commerce services. The simulation experiment shows that the proposed method can effectively improve the accuracy of the recommended system.

II. FORECASTING PRINCIPLE OF USER DEMAND

In the e-commerce service demand forecasting process, the concept of business service context ontology is divided into two types the concept of characteristics and the concept of e-commerce service characteristics according to the semantics, establishing the demand forecast rules. Current user environment and user requirements details are matched by the actual-time demand forecast reasoning of e-commerce. And forecasting the characteristics of user preferences to e-commerce services under the real-time situation, selecting service projects in line with the feature which will be recommended to the user from the feature set, and completing business services in the forecast. The concrete realization of the process as described below.

It is assumed that E represents the set of users and forecast objects, R^l is the concept of situation state, p^l is the concept of e-commerce service characteristics, U_i is the potential feature vector of e-commerce services, then user demand forecast rules are established by equation (1).

$$P_E = E \times \frac{R^l \times p^l}{U_i \times a_{\Re}} \times \mathrm{B}^{\chi} \times S_A \tag{1}$$

In equation (1), $a_{\Re}$ is the relationship network between users, B^{χ} is any two forecast objects, S_A is the relationship between the recommended objects.

Suppose that e representing the user's current environment, using the equation (2) to match the user current environment and the rules of the user requirements forecast.

$$T_y = \frac{e}{P_E} \times p^l \times S_A \tag{2}$$

Suppose that $q_{\Im}$ represents the trusted neighbor user group, I_j is on behalf of the similarity of user j with all neighbor candidate users, by using the equation (3) to forecast the user preferences to e-commerce service features under real-time situations, establish the objective function of user demand forecast in e-commerce requirement service, and complete user demand forecast.

$$P_{\xi} = \frac{e}{T_y \cdot a_{\Re} \times P_E} \tag{3}$$

But by the traditional method the recommendation can only be selected on the basis of single index, which reduces the accuracy of the demand forecast of e-

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00009

commerce service. A method of equilibrium forecast for e-commerce service demand based on ontology is proposed.

III. EQUILIBRIUM PREDICTION OF CUSTOMER DEMAND FOR E-COMMERCE SERVICE BASED ON ONTOLOGY

A. Calculation of similarity threshold between e-commerce users

When we do equilibrium prediction of customer demand for e-commerce service, we need to build user ontology for e-commerce service, getting the attribute system of user ontology knowledge, based on that system user attributes are divided into 3 different types. Firstly, when the users register, they provide the basic information of the document which is defined as user basic information. Secondly, it is the document of social information set, which we use to structure the users' knowledge. Thirdly, it is the information document of user's project set. If any user is interested in e-commerce service, he will evaluate the project that he is interested in. Accordingly, we define the document as the scoring information matrix of user interest items. According to the different types of the above documents, the basic attributes, main features and interest of a user can be described in details.

According to the principles above, the similarity of the two users' ontology is calculated by using the equation (4).

$$Sim_{property}\left(o_1,o_2\right)=\frac{\left|p_1\cap p_2\right|}{\left|p_1\cap p_2\right|+\theta\left\|p_1-p_2\right\|+\left|p_2-p_1\right|} \quad (4)$$

In equation (4), p_1 and p_2 mean the attribute set of O_1 and O_2 respectively. $\left|p_1\cap p_2\right|$ is the elements number of p_1 and p_2 intersection. $\left\|p_1-p_2\right\|$ is the elements number of set p_1 not belonging to set p_2. While $\left\|p_2-p_1\right\|$ represents the elements number of set p_2 not belonging to set p_1. θ represents the similarity threshold calculated by equation (5).

$$\theta=\frac{\left\{depth\left(O_1\right),depth\left(O_2\right)\right\}}{depth\left(O_1\right)+depth\left(O_2\right)\times Sim_{property}\left(o_1,o_2\right)} \quad (5)$$

In equation (5), $depth\left(O_i\right)$ is the maximum value of the entire conceptual path of ontology.

B. The accurate calculation of similarity between users

In the two users' ontology, there is the common concept of sequence. In the common sequence between ontologies the concept point is consistent, and the order is the same, which can reflect the similarity more accurately between two users' ontology. These 3 factors: the number of common sub sequence, levels, program are calculated respectively, then in a comprehensive summary, calculating the accurate similarity between the user ontology. Specific steps as follows.

1) The calculation of the number of common sub sequences of user ontology

It is assumed that $CS\left(O_1,O_2\right)$ represents the common sub sequence of two users' ontology. And the similarity of $CSnum\left(O_1,O_2\right)$ represents the common sub sequence is calculated by the equation (6).

$$CSnum\left(O_1,O_2\right)=\frac{2\times NSC\left(O_1,O_2\right)\times SC\left(O_1,O_2\right)}{S\left(O_i\right)+S\left(O_i\right)} \quad (6)$$

In equation (6), O_i is the number of its sub sequences, $NCS\left(O_1,O_2\right)$ represents the common sub sequences of the two ontology.

2) The calculation of the hierarchy of the common sub sequence of the user's ontology

Suppose that DC represents the depth of the common sub sequence, the similarity between the levels of the common sub sequence is obtained by the equation (7).

$$CS_{Pec}\left(O_1,O_2\right)=\frac{D_c-\theta\cdot depth\left(LCS\left(O_1,O_2\right)\right)}{CSnum\left(O_1,O_2\right)} \quad (7)$$

In formula (7), depth represents the distance from the concept to the root node of the ontology. $LCS\left(O_1,O_2\right)$ represents the minimum common sub sequence.

3) The length calculation of the common sub sequence of the user's ontology

In the two ontology, the longer the length of the common sub sequence is, the higher the similarity is between the corresponding two users. The length factor of the common sub sequence is obtained by using the equation (8).

$$CSlen\left(O_1,O_2\right)=\frac{Max\left(lenth\left(CS\left(O_1,O_2\right)\right)\right)}{CS_{Pec}\left(O_1,O_2\right)} \quad (8)$$

Combining the equations (6), (7) and (8), the similarity function of the user ontology is obtained by using the equation (9).

$$\mathrm{si}m\left(O_1,O_2\right)=\frac{f_1\times CSnum\left(O_1,O_2\right)CSlen\left(O_1,O_2\right)\bullet f_2}{CS_{Pec}\left(O_1,O_2\right)\otimes CSnum\left(O_1,O_2\right)} \quad (9)$$

In the equation (9), f_1 and f_2 represent the weights of the similarity between the two types of user ontology.

The initial weights of user ontology similarity function are set up in sigmoid. The similarity threshold above is taken into consideration, and the similarity function is introduced into the equation (9). The similarity between users is accurately calculated by using equation (10).

$$f\left(x\right)=\frac{1-e^{\varphi x}}{1+e^{\varphi x}}\times\frac{2}{1+e^{\varphi x}}\times\frac{\mathrm{si}m\left(O_1,O_2\right)}{\theta} \quad (10)$$

In equation (10), φ represents the fixed constant on the basis of specific application. x is the similarity value calculated by various factors and $f\left(x\right)$ is the initial weight of different factors.

C. Realization of equilibrium prediction of e-commerce service

On the basis of the user' ontology similarity, it collects the neighbor objects of the forecast targeted users,

establish groups of users' forecast demands, calculates the trusty subgroup with higher prediction probability, predicts the dynamic measurement of uncertain neighbors.

It effectively completes the prediction of users' demands. The specific implementation process, as shown below.

Assuming I_j representing user a of a service without scoring, $R_{a,j}$ is the forecast user scoring to the service. In the service without user's score, the forecast user provides the maximum score to recommended selectively to the user, by using the equation (10) to get $f(x)$ and search k neighbors for the unknown I_j, using the equation (11) to obtain the scores with user ontology forecast.

$$R_{a,j} = \bar{R}_x + \frac{sim'(U_a, U_b)^a \times (R_{b,j} - \bar{R}_b)^\delta}{\sum sim'(U_a, U_b) \times f(x) \cdot I_j} \quad (11)$$

In equation (10), $\bar{R}_a$ represents the scoring for other service types by user U_a in e-commerce, $sim'(U_a, U_b)$ representing the scoring intersection of user U_a and user U_b. $\bar{R}_b$ represents the scoring for other service types by user U_b.

It is assumed that $S(U_a)$ the prediction set selected by user ontology and $S(I_j)$ the prediction set selected by e-commerce, according to the equation (12), through them we calculate the common scoring number between user ontology more than user demand forecast group of the set threshold ε.

$$S'(U_a) = \varepsilon \cdot \frac{\{U_x \mid sim'(U_a, U_x) \geq \mu\}}{S(I_j) \times R_{a,j} \times S(U_a)} \quad (12)$$

In equation (12), U_x represents x users set. μ represents the threshold of similarity calculation between the service types of the e-commerce system.

By the equation (13), the trusty subgroup with higher prediction accuracy is obtained from the user's demand prediction group.

$$S'(I_j) = \varepsilon \cdot \frac{\{U_x \mid S'(U_a) \geq \mu\}}{S(I_j) \times R_{a,j} \times S(\gamma)} \quad (13)$$

In equation (13), γ is the set trusty degree threshold. By equation (14), the user demand equilibrium prediction model in the process of e-commerce service is established.

$$R'_{a,j} = \lambda \left[\bar{R}_a + \frac{(R_{x,j} - \bar{R}_x)}{sim'(U_a, U_b)} \times \frac{U_x \times (I_j, I_y)}{(\bar{R}_y, \bar{R}_j)} \right] \quad (14)$$

In equation (14), $\bar{R}_a$ and $\bar{R}_x$ represents the average score to other services of user U_a and U_x respectively. $\bar{R}_j$ and $\bar{R}_y$ represents the average score value of the total number of known users I_j and I_y. λ represents the neighbor of the user.

IV. Simulation Experiment Results and Analysis

In order to prove the effectiveness of the proposed method of user demand equilibrium prediction in the process of ontology e-commerce service, simulation experiments are needed. The data set used in the experiment is derived from the data of a university mobile catering service platform. During the experiment we totally recruited 200 student volunteers to participate in the use of the prototype system. The experimental time was 100 days. During the experiment, the experimenter had to order every meal through the interactive interface of mobile client inputting that time situation and and selecting restaurants. The system generated one user demand forecast table based on the situation.

A. Ratio of customer demand forecast efficiency in e-commerce service

By using the improved algorithm and literature 3, 4 and 5 algorithm to predict user demand in e-commerce service, the prediction efficiency of different methods is compared at different time, and the results are shown in Figure 1 as below.

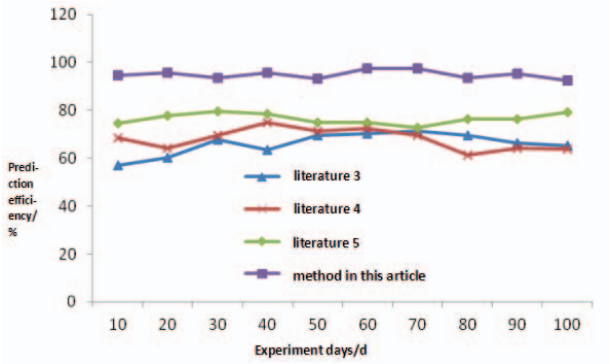


Figure 1. Comparison of the prediction efficiency of different algorithms.

Figure 1 analysis shows that with the same experimental days in the literature of 3, 4, and 5 the prediction efficiency all was lower than that of the improved algorithm. Through the analysis of the whole experimental process, we could know that the prediction average efficiency of literature 3, 4, and 5 is about 66.03%, 67.98%, 76.48% respectively, while that of the improved algorithm is about 94.92%, being increased by 43.73%, 39.62%, 24.10% to compare the three method above, which shows that the improved algorithm has high prediction efficiency.

B. The average absolute deviation ratio of user demand forecast in e-commerce service

Mean Absolute Error(MEA) which represents the average absolute deviation is defined as the recommended accuracy measurement standard for different algorithms. By MEA we calculate the deviation between the predicted user scoring and the actual user scoring. The smaller the MAE value is, the higher the accuracy of the recommendation results is. The equation (15) is used to calculate MAE.

$$MAE = \frac{1}{n}\left[r_{u,i} - \hat{r}_{u,i} \right] \quad (15)$$

In equation (15), $\hat{r}_{u,i}$ represents the forecast results, $r_{u,i}$ representing the collection of the types of e-commerce services, and n representing the size.

By the improved algorithm and literature 3, 4, and 5 algorithms to predict user demand in e-commerce service, the average absolute deviation prediction of different methods is compared at different time, and the results are shown in Figure 2 as below.

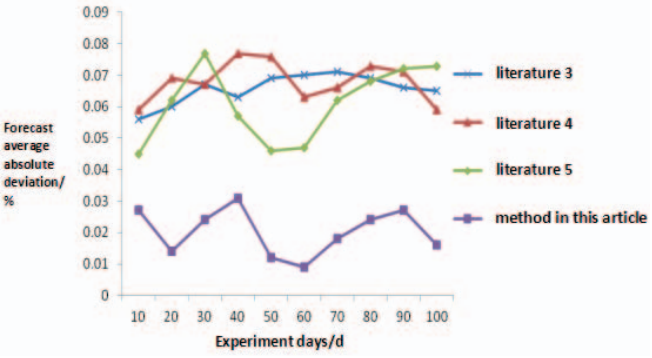


Figure 2. Comparison of predicted average absolute deviation of different algorithms.

Figure 2 shows that with the same experimental days under the condition that the average absolute deviation of literature 3, 4, and 5 is much higher than that of the improved algorithm. From the analysis of the whole experimental process, the prediction average efficiency of literature 3, 4, and 5 is about 0.0656%, 0.068% and 0.0609% respectively, while that of improved algorithm about 0.0202%, which shows that the improved algorithm has small prediction deviation.

C. Comprehensive and effective comparison of user demand forecast

By the improved algorithm and the algorithms of literature 3, 4, and 5 to forecast users' demands, within a given time period, comparing the accuracy of several algorithms (T)%, reliability E (%) and memory, with the results we could measure the comprehensive effectiveness of different algorithms to e-commerce service equilibrium recommendation. The results are shown in table 1.

TABLE I. THE COMPREHENSIVE EFFECTIVENESS WITH DIFFERENT ALGORITHMS TO USER DEMAND FORECAST OF E-COMMERCE SERVICE

Algorithms	(T)%	E（%）
algorithm of literature 3	69.17	62.91
algorithm of literature 4	71.13	65.28
algorithm of literature 5	76.31	73.12
Improved algorithms	95.32	94.38

Analysis of Table 1 shows that in terms of accuracy, the prediction accuracy of literature 3, 4, and 5 is about 69.17%, 71.13% and 76.31% respectively, while the improved method is about 95.32%. In terms of the reliability of prediction, the prediction accuracy of 3, 4, and 5 is about 62.91%, 65.28% and 73.12% respectively, while the improved method is about 94.38% . It is improved that this method is more effective than other three, and it can meet the actual demands of users.

ACKNOWLEDGMENT

This research was supported by the Application Foundation Frontier Project (2020010601012294), the Wuhan University of Business Doctoral Fund projects (2021KB002),the Teaching and Research Program at City-owned Universities in Wuhan (2019102), the Provincial University Teaching Research Project of Hubei Province (2018466), and the Construction of Specialty Features of Sports Economics and Management Major in Wuhan Business University (2019N013).

REFERENCES

[1] L. Guo, J. Ma, Z. M. Chen and H. R. Jiang, Incorporating Item Relations for Social Recommendation, Chinese Journal of Computers. 37 (2014) 219-228.

[2] A. T. Yang, Y. Tang, J. B. Wang and J. G. Li, Personalized Friends Recommendation System Based on Game Theory in Social Network, Computer Science. 42 (2015) 191-194.

[3] Z. H. Huang, Top-n Recommendation Algorithms for Cloud Data, Acta Electronica Sinica. 43 (2015) 54-61.

[4] Y. Jiang, Research and Simulation on Information Trust Calculation Model of P2P Network Nodes, Computer Simulation. 31 (2014) 298-301.

[5] J. Y. Tang, S. C. Huang, Application of optimized similarity calculation in recommender system, Electronic Design Engineering. 23 (2015) 46-48.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

A review on the issues in usability evaluation of eHealth applications

From a COVID-19 pandemic perspective

Bashair AlThani
College of Business Adminstration
Imam Abdulrahman Bin Faisal University
Dammam, Saudi Arabia
bfalthani@iau.edu.sa

***Abstract*— The use of eHealth applications has increased significantly across the globe due to the situations arising out of the recent COVID-19 pandemic. Many new eHealth applications such as mHealth applications for online consultations, health tracking and monitoring, diagnosis etc. were developed, which might have been evaluated using the existing techniques. However, the pandemic has created a situation, which can significantly influence the various processes adopted in the evaluation methods, which may prove to be ineffective due to the unexpected situations arising out of the COVID-19 pandemic. Therefore, the aim of this study is to review the various issues in usability evaluation of eHealth applications in the context of COVID-19 pandemic. The review has identified the need for emphasis on human factors such as attitudes, behaviors, lifestyles, knowledge, and skills; and other environmental factors such as competition, standards, regulations etc. was identified to be necessary in addition to the focus on technical aspects of the system in usability evaluation of eHealth systems.**

Keywords-eHealth, usability, evaluation, issues, applications

I. INTRODUCTION

Although, there is no clear consensus on what exactly comprises in eHealth, it is widely acknowledged that all healthcare operations which uses the Internet and communication technologies (ICTs) are considered as eHealth [1]. The recent COVID-19 pandemic has significantly influenced the adoption of eHealth technologies across the globe due to an increased burden on the healthcare systems. The total number of COVID-19 cases globally reached 237.3 million including 4.8 million deaths as reported on 12th October 2021 [2]. Since its emergence, the world has seen three COVID-19 waves at their peak in December 2020, April 2021, and August 2021, which have also seen significant rise in number of deaths [2]. The rise in the COVID-19 cases has forced the countries to implement various suppression and prevention strategies such as lockdowns, curfews etc., which affected the patients suffering from chronic illnesses and those with regular diseases, and the hospitals were already overcrowded with COVID patients. To address the issues in the delivery of healthcare services, various governments have relied on the increased adoption of eHealth technologies. As a result, various studies focused on the relevant aspects such as adoption of eHealth or digital health technologies [3], awareness or knowledge of computer skills, mobile applications [4], awareness of COVID-19 through eHealth applications [5] etc. Accordingly, the global funding for eHealth technology reached $13.9 billion in 2020; and it is expected to reach $38 billion in 2025 [6], indicating a significant growth in the investments in eHealth. According to Mckinsey report [7], the use of telehealth services globally, has increased by almost 80 times in volume in April 2020, which later remained stable at 38 times in volume before the pandemic.

Furthermore, the report estimated $250 billion spent on remote healthcare services in the US alone during the pandemic, as the consumer attitudes towards eHealth technologies have improved since the emergence of the pandemic, while the barriers such as security and privacy still remained. This indicates that the consumers may be using the eHealth technologies out of need, but not out of interest, as the months in the second half of the pandemic reflected a stable 38 times in volumes, which were dropped from 78 times in volumes in the beginning of the pandemic. Furthermore, there is an increase in the use of virtual agents or automated computer program or bots for providing remote healthcare services, and a significant growth in innovation of remote healthcare technologies was identified with the regulatory changes to promote remote healthcare access [7]. These recent changes: the sudden increase in the adoption of eHealth technologies and a surge in the innovation and development of effective and efficient eHealth technologies has increased the complexity of the usability evaluation of eHealth applications, as the new dimensions and perceptions were emerged in the process of evaluation. The evaluation process of eHealth systems is no more considered as a piece of software or hardware, or a set of interconnected devices, but also consider the actors or stakeholders, physical spaces, and policies. It has become essential to consider the attitudes, behaviors, and the perceptions of the consumers in evaluating the usability of various eHealth applications. For instance, evaluating the COVID-19 monitoring eHealth system may need to consider various perspectives such as the perspectives and conditions of people with issues such as mental disorders, anxiety etc.

As a result, there is a possibility of emergence of new issues in the usability evaluation of eHealth. In this context, this paper reviews the issues in usability evaluation of eHealth applications in different categories presented in the following sections.

A. Technical reliability and appropriateness

While reliability is not a direct non-functional user requirement, but customer requirements should be expressed in terms of benefits reliability can deliver. Compared to other systems, the risk of failure or an error in eHealth system can cause severe damage, sometimes may result in death of a patient. With the increase in the technology interventions, the scope of eHealth has

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00010

increased to a wide extent. The eHealth has emerged from the use of single services such as monitoring and tracking of glucose levels through Bluetooth enabled glucose device and smartphone to 24x7 monitoring of glucose, and various other health indicators such as heart rate, blood pressure, oxygen levels, and also behavioral aspects such as sleep time, stress etc., using interconnected devices such as sensors, wearables, and many other interconnected devices, typically referred as the Internet of things (IOT) [8]. Evaluating the reliability and appropriateness of IOT eHealth applications can be challenging, as there are various aspects that need to be evaluated such as accuracy of information (health vitals) being recorded, appropriate communication between the devices etc. For instance, if there is an issue in transferring the glucose data/oxygen levels of an elderly COVID-19 patient by continuous glucose monitor to a connected IOT device, it may cause serious damage to her organs and may sometimes even lead to death. Therefore, not only the device, but also the networks, the communicative environments, protocols etc. must be evaluated. However, in the context of COVID-19, as it is still being studied, there are only few vital measures that are being considered in monitoring patients, such as heartbeat, pulse, oxygen levels etc. However, with the emergence of research, various other measures or indicators may be introduced, and integrating and evaluating these in the new eHealth models may be challenging. Therefore, with the speedy research related to COVID-19, many new measures, indicators, policies, standards, and regulations are being introduced; which makes it a challenging task to evaluate the usability of eHealth applications in the agile environments.

B. Privacy & Security

Privacy is still one of the major issues affecting the technology industry. Though there were significant advances in developing the solutions to protect privacy, breaches are still being observed every now and then in different businesses. The main concern of eHealth systems is the protection of patients' data. Electronic Medical Records (EMRs) or Electronic Health Records (EHRs) have various benefits to both patients and healthcare providers, as they provide complete health related information about patients, including the critical information such as diagnosis reports, medication, personal details etc. However, ensuring the patients' privacy by protecting EHRs is one of the most challenging tasks in managing eHealth systems [9]. As a part of automation of healthcare services, various countries are moving towards adoption of integrated eHealth systems and developing EHRs; but at the same time, the number of security incidents have been increasing. For instance, Anthem, a health insurance provider lost 80 million records of its customer due to a hacker attack, which forced the company to pay $39.5 million as a part of settlement with the States in the US [10]. Similarly, Science Applications International Corporation (SAIC), was also affected by the data breach resulting in the loss of 4.9 million military clinic and patients records in the US [11].

During the COVID-19, there is a sudden surge in the eHealth users, which can make the existing eHealth systems face various issues in managing the data; which can become one of the advantageous situations for cyber attackers [12]. According to Deloitte report, the cyber-attacks during COVID-19 have increased by 35% compared to non-pandemic situation [13] in Switzerland. Therefore, it is evident that COVID-19 has affected the usability evaluation of eHealth applications due to sudden change in the number of users and emergency updates in system environment and architecture in various eHealth systems.

C. Interoperability

Interoperability is one of the major challenges in eHealth due to the large volumes of data such as EHRs, EMRs, patient health records, and other types of data such as diagnostic data, monitoring data through wireless/wired sensors, and varying levels of standards and regulations etc. The complexity in the information to be managed is one of the major challenges faced by the eHealth technology developers and the companies adopting eHealth technologies. Furthermore, different types of information which needs to be recorded and managed through the lifetime of the patients, during which there may be several challenges incurred due to the upgradation of legacy systems, implementation/integration of new technology solutions to ramp up the process efficiency and cost reduction and achieve competitive advantage [14-16]. For example, adopting cloud-based technology into eHealth infrastructure may involve serious challenges in integrating data, maintaining the quality of data, and ensuring the successful integration. With rapid changes in the eHealth technologies, such as integration of IOT, which synchronizes the different types of data between various interconnected devices may further create complexity in managing interoperability issues due to the integration of existing systems, i.e., non-interoperable IOT systems into interoperable IOT systems [17-18].

During the COVID-19, sudden surge in the number of eHealth users could have increased the need for additional eHealth architecture, and integration of various systems and data types, which can have serious impact on the functionality, reliability, and accessibility [19,20]. Lack of effective global standards in eHealth management is considered to be one of the major reasons for the issues associated with interoperability in eHealth technologies [21]. Therefore, the future research may focus on developing the global eHealth standards.

D. Gig economy/ outsourcing

Gig economy is associated with the process of outsourcing gigs (small works) to independent gig workers rather than employing permanent employees by the organizations, which can significantly reduce operational costs. Gig economy has a promising future in eHealth as it can benefit the process of healthcare management more efficiently and effectively [22]. Furthermore, gig economy can also benefit in reducing the operational costs by outsourcing gig works [23]. For instance, in eHealth technologies; health service providing mobile applications can provide gigs such as online consultations to professional physicians; and also, other gigs such as online medicine ordering and deliver; or diagnostic gigs such as collection of blood samples for tests at home. Outsourcing such operations to gig workers reduces the costs, as the payment is based on the gig works but not on monthly

basis. However, there may be some serious issues in this process such as privacy and security, as healthcare data is being accessed by the third-party physician/nurse. Furthermore, integrating the patients' health records managed by third-party mobile application/consultant may involve challenges such as data integrity, interoperability issues etc. There was a significant rise in the downloads of mobile health applications which is identified to be 65% growth during COVID-19 pandemic early days compared to prior COVID-19 pandemic period [24]. Furthermore, there is an increasing interest among the physicians to join the gig workforce as it offers flexibility in work, as required [25]. With the healthcare industry, experiencing new business models such as gig economy, the usability evaluation of the eHealth technologies is becoming more challenging due to the changes in business models and users' attitudes, behaviors, requirements, and most importantly expectations.

E. Awareness, acceptance, and readiness

Awareness, adoption, and readiness of the eHealth consumers are the important factors that need to be considered in the usability evaluation of eHealth application. Awareness of eHealth applications such as mHealth (how to use/register, booking appointments etc.), acceptance (acceptance of approaches such as online consultations, which may not go well with certain cultures), and readiness of the consumers for using eHealth applications (possessing enough knowledge and skills) are the three important factors that need to be considered in eHealth evaluation in order to achieve success in its implementation. Studies [26-29] have shown varying levels of eHealth adoption, which is considerably high in high-income countries compared to low-income countries; lack of effective standards or policies in few countries; varying levels of skills and awareness about eHealth among the public; and most importantly lack of regular evaluation of eHealth applications. These differences may increase the complexity of usability evaluation. For instance, in regions where people have less computer/mobile usage skills, the usability evaluation has to be done according to their knowledge levels but not according to the testers' knowledge in organizations. Furthermore, the evaluation process has to be changed according to the changes in the awareness and knowledge of the public with time, which requires regular usability evaluation of eHealth applications. During the COVID-19 pandemic, the rise in the number of eHealth subscribers reflect a sudden change in attitudes of the people, which doesn't mean that they suddenly became aware of eHealth and possess sufficient skills for its adoption. It is possible that their apps are operated by others such as friends/relatives for accessing services such as booking vaccines, booking appointments etc., which can be commonly observed in developing and under-developing countries. Therefore, ever-changing behaviors, lifestyles, knowledge, and awareness of the eHealth users can increase the complexity and challenges in usability evaluation of eHealth applications.

F. Information management

Information overload is one of the major issues affecting the implementation of ICTs in all sectors including, health, education, business etc. Unlike other sectors, healthcare is one of the most important sectors in which large volumes of information is produced every day, which has to be managed effective for many years. Many issues may emerge in this process, which may include filter failures such as inadequate information retrieval systems for point-of-care settings, the problem of identifying all relevant evidence in an exceedingly diverse landscape of information resources, and the very basic lack of health information literacy [30]. The process of finding solutions to address the problem of information management largely concentrated on technological means while undermining the humans/patients in the process of managing information [31]. In addition, information overload can also affect the healthcare services such as emergency consultations, in which physicians may need to assess large volumes of patients' information in short times, which not only affects the patients' treatment but also affects physician's health by increasing anxiety and stress [32,33]. Similar cases with information overload leading to anxiety and stress due to information overload was observed among the elderly users of mHealth applications in a study conducted in China [34].

Information overload, such as spread of myths or false information during the pandemic has created a lot of havoc such as burning of 5G towers in the UK, drinking alcohol in Iran for curing COVID-19 etc. Addressing these myths by creating awareness through eHealth applications is a challenging process that increases the real-time monitoring on various issues, and accordingly updating the notifications and messages to the users [35]. Furthermore, evaluating the eHealth applications from the perspectives of people affected by these myths can be challenging, as they may either influence other people over the eHealth network through posts. Therefore, the need for continuous evaluation of eHealth applications may be necessary in order to prevent any damage from the people influenced by false information.

G. Malpractices

Malpractices in eHealth have been increasing in the past few years. Especially the EHR-related claims were on the rise, which were mostly associated with user/physician related aspects such as sloppy copy-and-paste habits, alert fatigue, and workarounds; and also related system functionalities such as data routing problems, inappropriate drop-down menus and failed clinical decision support software [36]. These malpractices reflect a serious patients safety risks associated with the use of EHRs, one of the main components of eHealth system. However, few studies [37] have found that there is a significant decrease in the malpractices after the implementation of EHRs. Similarly, in a study [38] conducted on 1884 physicians in Massachusetts, it was identified that there were 6.1% of physicians with an EHR had a history of a paid malpractice claim compared with 10.8% of physicians without EHRs; indicating the considerable existence of malpractices even after implementing the EHRs. These issues raise the concerns such as liability as the patients' privacy and security are at stake, and the studies [39,40] have highlighted the need for effective training and implementing standards of practice; and the need for physicians to adapt to efficient and effective use of the

electronic information highway to address the issues of malpractices.

Cases of malpractices in eHealth were seen to be increasing during the COVID-19 pandemic, which may be attributed to various reasons such as lack of training and support for physicians, lack of complete knowledge about the treatment procedures for COVID-19, ambiguity in the standards and protocols etc. [41,42]. Therefore, in the event of sudden change such as COVID-19, though it is common to identify the malpractices, it is highly important that the systems are to be effectively evaluated along with the human (physicians/frontline workers) behaviors and attitudes in order to determine the reliability and usability of the eHealth systems. Furthermore, these issues such as malpractices or system errors may lead to poor or ineffective decision-making which may result in severe risks associated with patients' safety. Therefore, evaluating the eHealth system for such issues as malpractices associated with humans may be challenging, as the tester may need to adopt psychology of human behaviors while evaluating the systems.

II. Conclusion

COVID-19 pandemic has significantly influenced every aspect of life and business, which have led to the rise of significant challenges in every field of business and service. Healthcare/eHealth is also an important sector that has been influenced by the pandemic. Significant changes have been observed in eHealth in the areas of resource utilization and management; information systems development, implementation, and maintenance; operations and procedures; standards of practice etc. However, the changes in all these aspects have significantly influenced the process of usability evaluation of eHealth applications, which required the reengineering of various procedures and methods of evaluation, and most importantly considering the human related factors such as attitudes, behaviors, knowledge and skills etc.

It is evident from the review that COVID-19 has significantly affected usability evaluation of eHealth applications; and the identification of various areas in which such issues are experienced is an important contribution of this study, which has both theoretical and practical implications. The findings of this review support the need for extending the theory of usability evaluation of eHealth systems in emergency or unexpected situations such as COVID-19 pandemic; and also has practical implication, as the findings can be used as a source of information by software or system testers or evaluators for evaluating the usability of eHealth applications. This study also has few limitations, as it only considered seven important areas, where the issues in usability evaluation of eHealth systems was considered; but there can be more areas to be investigated such as trust, decision-making, liability, legal regulations etc., which may be addressed in future studies.

References

[1] H. Oh, A. Jadad, C. Rizo, M. Enkin, J. Powell and C. Pagliari, "What Is eHealth (3): A Systematic Review of Published Definitions", Journal of Medical Internet Research, vol. 7, no. 1, 2005. Available: 10.2196/jmir.7.1.e1 [Accessed 12 October 2021].

[2] World health organization, "WHO Coronavirus (COVID-19) Dashboard", COVID19.who.int, 2021. [Online]. Available: https://COVID19.who.int/. [Accessed: 12- Oct- 2021].

[3] Hamza, M.S., Badary, O.A. & Elmazar, M.M. Cross-Sectional Study on Awareness and Knowledge of COVID-19 Among Senior pharmacy Students. J Community Health Volume 46, 139–146, (2021). https://doi.org/10.1007/s10900-020-00859-z

[4] Bokolo, A.J. Application of telemedicine and eHealth technology for clinical services in response to COVID 19 pandemic. Health Technol, Volume 11, 359–366 (2021). https://doi.org/10.1007/s12553-020-00516-4

[5] X. Li and Q. Liu, "Social Media Use, eHealth Literacy, Disease Knowledge, and Preventive Behaviors in the COVID-19 Pandemic: Cross-Sectional Study on Chinese Netizens", Journal of Medical Internet Research, vol. 22, no. 10, p. e19684, 2020. Available: 10.2196/19684 [Accessed 12 October 2021].

[6] M. Mikulic, "Topic: Digital health", Statista, 2021. [Online]. Available: https://www.statista.com/topics/2409/digital-health/. [Accessed: 12- Oct- 2021].

[7] O. Bestsennyy, G. Gilbert, A. Harris, and J. Rost. "Telehealth: A quarter-trillion-dollar post-COVID-19 reality?" [online] Available: https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/telehealth-a-quarter-trillion-dollar-post-COVID-19-reality [Accessed: 12- Oct- 2021].

[8] C. Butpheng, K.-H. Yeh, and H. Xiong, "Security and Privacy in IoT-Cloud-Based e-Health Systems—A Comprehensive Review," Symmetry, vol. 12, no. 7, p. 1191, Jul. 2020.

[9] I. Keshta and A. Odeh, "Security and privacy of electronic health records: Concerns and challenges", Egyptian Informatics Journal, vol. 22, no. 2, pp. 177-183, 2021. Available: 10.1016/j.eij.2020.07.003 [Accessed 12 October 2021].

[10] R. Staff, "Anthem to pay nearly $40 million to settle data breach probe by U.S. states", U.S., 2021. [Online]. Available: https://www.reuters.com/article/us-anthem-cyber-idUSKBN26L2PW. [Accessed: 12- Oct- 2021].

[11] J. Forsyth, "Records of 4.9 mln stolen from car in Texas data breach", U.S., 2021. [Online]. Available: https://www.reuters.com/article/us-data-breach-texas-idUSTRE78S5JG20110929. [Accessed: 12- Oct- 2021].

[12] Muthuppalaniappan, Menaka, and Kerrie Stevenson. "Healthcare cyber-attacks and the COVID-19 pandemic: an urgent threat to global health." International journal for quality in health care : journal of the International Society for Quality in Health Care vol. 33,1 (2021): mzaa117. doi:10.1093/intqhc/mzaa117

[13] Deloitte, "Impact of COVID-19 on Cybersecurity", Deloitte Switzerland, 2021. [Online]. Available: https://www2.deloitte.com/ch/en/pages/risk/articles/impact-COVID-cybersecurity.html. [Accessed: 12- Oct- 2021].

[14] A. Dogac, "Interoperability in eHealth systems", Proceedings of the VLDB Endowment, vol. 5, no. 12, pp. 2026-2027, 2012. Available: 10.14778/2367502.2367568 [Accessed 13 October 2021].

[15] C. Kuziemsky, N. Archer and L. Peyton, "Towards E-Health Interoperability: Challenges, Perspectives and Solutions", Journal of Emerging Technologies in Web Intelligence, vol. 1, no. 2, 2009. Available: 10.4304/jetwi.1.2.107-109 [Accessed 13 October 2021].

[16] B. Nathan et al. "Data Quality and Interoperability Challenges for eHealth Exchange Participants: Observations from the Department of Veterans Affairs' Virtual Lifetime Electronic Record Health Pilot Phase." AMIA ... Annual Symposium proceedings. AMIA Symposium vol. 2014 307-14. 14 Nov. 2014

[17] Behr Technologies, "Interoperability: The Secret to a Scalable IoT Network", BehrTech, 2021. [Online]. Available: https://behrtech.com/blog/interoperability-the-secret-to-a-scalable-iot-network/. [Accessed: 13- Oct- 2021].

[18] M. Noura, M. Atiquzzaman and M. Gaedke, "Interoperability in Internet of Things: Taxonomies and Open Challenges", Mobile Networks and Applications, vol. 24, no. 3, pp. 796-809, 2018. Available: 10.1007/s11036-018-1089-9 [Accessed 13 October 2021].

[19] R. Crutzen, "And Justice for All? There Is More to the Interoperability of Contact Tracing Apps Than Legal Barriers.

Comment on “COVID-19 Contact Tracing Apps: A Technologic Tower of Babel and the Gap for International Pandemic Control”", JMIR mHealth and uHealth, vol. 9, no. 5, p. e26218, 2021. Available: 10.2196/26218 [Accessed 13 October 2021].

[20] M. Krausz, J. Westenberg, D. Vigo, R. Spence and D. Ramsey, "Emergency Response to COVID-19 in Canada: Platform Development and Implementation for eHealth in Crisis Management", JMIR Public Health and Surveillance, vol. 6, no. 2, p. e18995, 2020. Available: 10.2196/18995 [Accessed 13 October 2021].

[21] ITU, "E-health Standards and Interoperability", Itu.int, 2021. [Online]. Available: https://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000170001PDFE.pdf. [Accessed: 13- Oct- 2021].

[22] T. Alanzi, "Gig Health vs eHealth: Future Prospects in Saudi Arabian Health-Care System", Journal of Multidisciplinary Healthcare, vol. 14, pp. 1945-1953, 2021. Available: 10.2147/jmdh.s304690 [Accessed 13 October 2021].

[23] A. Prestia, "Leveraging the Gig Economy: A Novel Solution to Improve Health Care Costs", Nurse Leader, vol. 17, no. 4, pp. 356-359, 2019. Available: 10.1016/j.mnl.2018.11.003 [Accessed 13 October 2021].

[24] C. Stewart, "COVID-19 growth in medical app downloads by country 2020 | Statista", Statista, 2021. [Online]. Available: https://www.statista.com/statistics/1181413/medical-app-downloads-growth-during-COVID-pandemic-by-country/. [Accessed: 13- Oct- 2021].

[25] P. Alperin, "PHYSICIANS AND THE GIG ECONOMY", Physicianleaders.org, 2021. [Online]. Available: https://www.physicianleaders.org/news/physicians-and-the-gig-economy. [Accessed: 13- Oct- 2021].

[26] S. Ryu, "Book Review: mHealth: New Horizons for Health through Mobile Technologies: Based on the Findings of the Second Global Survey on eHealth (Global Observatory for eHealth Series, Volume 3)", Healthcare Informatics Research, vol. 18, no. 3, p. 231, 2012. Available: 10.4258/hir.2012.18.3.231 [Accessed 13 October 2021].

[27] R. Bonyan, A. Al-Karasneh, F. El-Dahiyat and A. Jairoun, "Identification of the awareness level by the public of Arab countries toward COVID-19: cross-sectional study following an outbreak", Journal of Pharmaceutical Policy and Practice, vol. 13, no. 1, 2020. Available: 10.1186/s40545-020-00247-x [Accessed 13 October 2021].

[28] I. Hambarova & N. Mateva. PUBLIC AWARENESS ON EHEALTH IN BULGARIA - A PILOT STUDY. Knowledge International Journal, Vol. 35, no.4, pp. 1323 – 1328, 2019.

[29] M. O’Connor, K. McGowan and R. Jolivet, "An awareness-raising framework for global health networks: lessons learned from a qualitative case study in respectful maternity care", Reproductive Health, vol. 16, no. 1, 2019. Available: 10.1186/s12978-018-0662-9 [Accessed 13 October 2021].

[30] I. Klerings, A. Weinhandl and K. Thaler, "Information overload in healthcare: too much of a good thing?", Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, vol. 109, no. 4-5, pp. 285-290, 2015. Available: 10.1016/j.zefq.2015.06.005 [Accessed 13 October 2021].

[31] A. Hall and G. Walton, "Information overload within the health care system: a literature review", Health Information & Libraries Journal, vol. 21, no. 2, pp. 102-108, 2004. Available: 10.1111/j.1471-1842.2004.00506.x [Accessed 13 October 2021].

[32] L. Sbaffi, J. Walton, J. Blenkinsopp and G. Walton, "Information Overload in Emergency Medicine Physicians: A Multisite Case Study Exploring the Causes, Impact, and Solutions in Four North England National Health Service Trusts", Journal of Medical Internet Research, vol. 22, no. 7, p. e19126, 2020. Available: 10.2196/19126 [Accessed 13 October 2021].

[33] Z. Fritz, A. Schlindwein and A. Slowther, "Patient engagement or information overload: patient and physician views on sharing the medical record in the acute setting", Clinical Medicine, vol. 19, no. 5, pp. 386-391, 2019. Available: 10.7861/clinmed.2019-0079 [Accessed 13 October 2021].

[34] Y. Cao, J. Li, X. Qin and B. Hu, "Examining the Effect of Overload on the MHealth Application Resistance Behavior of Elderly Users: An SOR Perspective", International Journal of Environmental Research and Public Health, vol. 17, no. 18, p. 6658, 2020. Available: 10.3390/ijerph17186658 [Accessed 13 October 2021].

[35] Rathore, Farooq Azam, and Fareeha Farooq. “Information Overload and Infodemic in the COVID-19 Pandemic.” JPMA. The Journal of the Pakistan Medical Association vol. 70(Suppl 3),5 (2020): S162-S165. doi:10.5455/JPMA.38

[36] A. Allen, "Review finds increase in EHR-related malpractice claims", POLITICO, 2021. [Online]. Available: https://www.politico.com/tipsheets/morning-ehealth/2017/10/16/review-finds-increase-in-ehr-related-malpractice-claims-222821. [Accessed: 13- Oct- 2021].

[37] e. Network, "EHR adoption reduces malpractice claims - eHealth Magazine", Ehealth.eletsonline.com, 2021. [Online]. Available: https://ehealth.eletsonline.com/2012/06/ehr-adoption-reduces-malpractice-claims/. [Accessed: 13- Oct- 2021].

[38] A. Virapongse et al., "Electronic Health Records and Malpractice Claims in Office Practice", Archives of Internal Medicine, vol. 168, no. 21, p. 2362, 2008. Available: 10.1001/archinte.168.21.2362 [Accessed 13 October 2021].

[39] S. Hoffman. "E-Health Hazar E-Health Hazards: Provider Liability and Electr vider Liability and Electronic Health Recor onic Health Record Systems". Faculty Publications. 2, [online] Available at: https://scholarlycommons.law.case.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1001&context=faculty_publications [Accessed: 12 Oct 2021]

[40] Paterick, Zachary R et al. “Medical liability in the electronic medical records era.” Proceedings (Baylor University. Medical Center) vol. 31,4 558-561. 11 Sep. 2018, doi:10.1080/08998280.2018.1471899

[41] B. Clio et al. “The medico-legal implications in medical malpractice claims during COVID-19 pandemic: Increase or trend reversal?.” The Medico-legal journal vol. 88,1_suppl (2020): 35-37. doi:10.1177/0025817220926925

[42] S. Dash, R. Aarthy and V. Mohan, "Telemedicine during COVID-19 in India—a new policy and its challenges", Journal of Public Health Policy, vol. 42, no. 3, pp. 501-509, 2021. Available: 10.1057/s41271-021-00287-w [Accessed 13 October 2021].

Political economy analysis of carbon emissions

Based on empirical data from the central region

Lipeng Wen,Tingting Yin,Xuemei Long*
School of business
Guilin University of Electronic Technology
Guilin, People's Republic of China
1937259042@qq.com

KengCheng Zheng
School of Finance and Taxation
Zhongnan University of Economics and Law
Wuhan, People's Republic of China
1257992189@qq.com

***Abstract*—In order to correctly understand the influencing factors of carbon emissions in central China and the important role of government governance capacity in achieving the goal of "carbon neutrality". In theory, unit carbon emission is explained by Marxist axiology. In the empirical analysis, this paper selects the data from 2004 to 2017 in the central region as samples to verify the theoretical part, and finally puts forward relevant policy suggestions based on the above conclusions, so as to realize the sustainable economic growth in the central region.**

Keywords- Carbon dioxide emission; traditional factor; endowment effect; industrial; structure effect technological progress effect

I. INTRODUCTION

In the context of new economic development, global warming and greenhouse gas issues are once again widely concerned. From Marx to China's leaders for the unremitting efforts of ecological protection. Now, the 19th National Congress of the Communist Party of China has further put forward the strategic goal of building a "beautiful China" and made the construction of ecological civilization an important part of realizing the goal of building a modern socialist country. China's total carbon emissions account for about 25% of global carbon emissions, and as the world's largest carbon emitter [1]. The 19th National Congress of the Communist Party of China pointed out that "China's economic development model has been transformed into a high-quality development model, accelerating the pace of transformation from extensive growth to intensive growth, and taking the road of green development [2]." At the general debate of the 75th Session of the United Nations General Assembly held on September 22, 2020, China proposed to strive for the peak of carbon emissions by 2030 and achieve carbon neutrality by 2060.

On April 23, 2021, the Opinions on Promoting High-quality Development in the Central Region in the New Era was reviewed and approved by The State Council. The central region is an important carrying area for the development of The Yangtze River Economic Belt and the ecological protection and high-quality development of the Yellow River Basin in China, and the economic volume of the central region accounts for more than 1/5 of the national population and nearly 1/4 of the national population [3]. To promote high-quality development in the central region, energy consumption and carbon dioxide emissions in the six provinces of the central region have also increased sharply. So how can the six provinces in the central China achieve high-quality development while maintaining stability in the context of "carbon neutrality"?

At present, most domestic scholars have analyzed the relationship between carbon emissions and economic growth from the perspective of technology. Wang Jie (2021) discussed the relationship between carbon emissions and economic growth based on the carbon emission panel data of BRICS countries from 1987 to 2017, and used IPAT equation and LMDI model to decompose the driving factors affecting carbon emissions change [4]. Liu Manzhi (2021) developed decomposition and decoupling technology based on LMDI method in combination with C-D production function and quantified seven effects, providing a new perspective for energy conservation and emission reduction in transportation industry [5]. Although econometrics and technical analysis of the development of the relationship between the two often only exist in describing the characteristics of economic phenomena and operation mechanism, has a strong operability. Domestic research on the essential economic phenomenon between carbon emissions and economic growth is very scarce. Qi Xinyu (2010) used Marx's reproduction model and the basic idea of division of two major categories to conclude that carbon emissions can be reduced by slowing down the pace of economic growth, improving technological level and increasing the proportion of the second sector [6].

II. THE DECOMPOSITION OF REGIONAL CARBON EMISSION EFFECT

Analyzing the principle of ecological destruction from the perspective of Marx, first of all, the nature of capital is to pursue profit, and the surplus value created by laborers is the only source of profit. Capital expands reproduction and increases the proportion of surplus value production to obtain more surplus value. With the excessive demand, the discharge of waste in the process of absorbing surplus value gradually exceeds the carrying capacity of the ecosystem, leading to ecological damage. As Marx said, "Only by correctly understanding the laws of nature can human beings coexist harmoniously with nature, and the development and change of nature can make the situation change in the direction of human economic and social development [7]. Second, the capitalist mode of production as the main form of the economic growth of industrial civilization is invested capital, labor, land, natural resources), and technology so as to achieve the proliferation of capital and the increase of material wealth, as the pursuit of better add value, industrial civilization need to constantly in natural resources and bring a lot of cost and excessive use of natural resources. Finally, labor

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00011

originally changes natural material form with the help of nature, but with the improvement of labor productivity, labor appears "alienation". "The improvement of labor productivity and the increase of labor quantity come at the cost of the destruction and decline of labor force itself."

This paper tries to build a political economy model of carbon emissions, find the factors that affect carbon emissions, and discuss from the perspective of policy impact and government governance mechanism.

Assume that the total value of the industrial sector of a region in period t is w_t;In the process of value production, the total input constant capital value is c_t; The value of the variable capital invested in the production process is v_t; The wages of the workers are denoted by $wage_t*N_t$; The industry average residual value ratio is m', it can be considered as the average profit rate of the industry . This paper assumes that e_i is the emission coefficient of this region in phase t. Assume that Carbon Emission per Value (CEV) represents the Carbon dioxide emitted per unit of Value. The model is expressed as:

$$CEV_t = e_t c_t/[c_t + (m_t' + 1)v_t] \quad (1)$$

For simplicity, assume that the unit value of constant capital is $\bar{c}$ and the growth rate of constant capital is α_t; Variable capital v_t can be determined by the wages of the workers and the number of workers, that is, $v_t = wage_t N_t$. By introducing the organic composition of capital, the organic composition of capital ($s_t = \alpha_t \bar{C}/wage_t N_t$), we can obtain the following formula:

$$CEV_t = (e_t * \alpha_t \bar{c})/[wage_t N_t(1 + m' + s_t)] \quad (2)$$

Here, we divide the industrial sector into two sectors: high carbon industry and low carbon industry. It is assumed that the proportion of high carbon industry is θ^h and the proportion of low carbon industry is θ^l, and the proportion of low carbon industry and the proportion of high carbon industry are $\theta^h + \theta^l = 1$, $0 < \theta^h < 1, 0 < \theta^l < 1$. The emission intensity of the carbon industry is e^h, and that of the low carbon industry ise^l, and satisfies $e^h > e^l$, $e^l > e^h/m$.

Then the following formula can be obtained:

$$e_t \alpha_t \bar{c} = (\alpha_t e^h/m) * [(m-1)\theta^h + 1] \quad (3)$$

Government environmental control is mainly to take necessary administrative punishment for productive emissions to restrain enterprises, while government governance investment is to control existing environmental pollution emissions [8]. Therefore, we get the following formula:

$$e^h = f(regu, gov, rd) \quad (4)$$

Where, gov represents government governance investment and rd represents technology level. The above satisfies: $\partial e^h/\partial gov < 0$. Based on the economic substance, we derive the function expression of e^h basically as follows:

$$e^h = \beta/(gov \times rd) \quad (5)$$

Equation (3) can be further derived to obtain the following equation:

$$CEV_t = \frac{\alpha_t \times \beta \times \bar{c}[(m-1)\theta^h + 1]}{wage_t N_t(1 + m' + s_t) \times gov \times rd \times m} \quad (6)$$

From Equation (6), we can draw the following conclusions: ① The impact of factor endowment effect on carbon emissions per unit value is not clear, and the factor endowment effect is dynamically determined by the scale of capital investment, human capital investment and the organic composition of capital;② The relationship between government environmental regulation, government governance investment and technology level and the carbon dioxide emission per unit value; ③ The proportion of high-carbon industries is positively correlated with CO_2 emissions per unit value. If the development level of low-carbon and high-carbon industries in a region is unbalanced and the industrial structure is unreasonable, the CO_2 emissions in the region will rise.

III. DATA SOURCES AND EMPIRICAL ANALYSIS

The above assumptions of our political economy model are based on value theory. In order to verify our above theory, we use real data to conduct empirical analysis and test, in order to overcome the problem of inconsistency and disunity of measurement units. In order to find the internal relationship between value and price, Liu Xin (2000) revealed the internal logical contradiction of neoclassical general equilibrium theory and tried to construct Marx's general equilibrium theory model. Later, Feng Jinhua (2012) proved that there must be and only one general equilibrium price vector exactly equal to the corresponding value vector. Therefore, this part uses economic data under the price system to verify the feasibility of the carbon emission model under the value system [9].

1 Estimation of CARBON dioxide emissions (e)

This paper refers to the calculation method of carbon dioxide emissions released by IPCC, and the carbon dioxide emissions are estimated based on the energy consumption in central China, that is, the following formula is obtained to calculate carbon dioxide emissions:

$$CO_{2i} = \sum_{i=1}^{n} E_i \alpha_i \quad (7)$$

For the estimation of carbon dioxide emission coefficient, the following formula is used:

$$\alpha_i = NCV_i \bullet CEF_i \bullet COF_i \bullet (44/12) \quad (8)$$

We gradually arrive at the following formula:

$$CO_{2i} = \sum_{i=1}^{3} E_i \bullet NCV_i \bullet CEF_i \bullet COF_i \bullet (44/12) \quad (9)$$

The method of Zhao Xin (2021) is adopted for carbon oxidation factor. We calculate the carbon dioxide

emission coefficients for coal, oil and gas. Details are as follows:

TABLE I. CARBON DIOXIDE EMISSION COEFFICIENT

energy	*Average low calorific value*	*Carbon emission coefficient*	*Carbon oxidizing factor*	*Carbon dioxide emission coefficient estimates*
	(kj/kg)	*(kg / 10^6 kj)*	*/*	*(kg/kg standard coal)*
Raw coal	20908	26.0	0.99	2.763
Crude oil	41816	20.0	1	2.145
Natural gas	38931	15.3	1	1.642

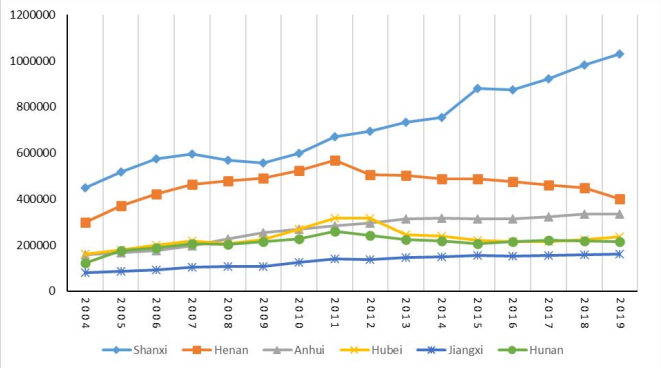


Figure 1. Carbon emission trend of the six central provinces.

2 Total output (y)

Total output is expressed in terms of gross domestic product (GDP) of the central region, which eliminates the effect of prices.

3 Capital stock per capita of the industrial sector

There is no direct data available for the per capita capital stock of the industrial sector in the statistical yearbook. However, due to the different degree of industrialization in each region, this paper adopts the accumulated depreciation amount of industrial enterprises above designated size to represent the factor of capital input [10].

4 Industrial structure (Indus)

The proportion of high-carbon industry is to indicate whether the industrial structure of a certain region is reasonable. This paper intends to use the ratio of the added value of the tertiary industry in each region to the added value of the secondary industry in each region to measure.

5 Average wage of industrial sector

In this paper, the average wage and employee number of mining industry, manufacturing industry and electricity, gas and water production and supply industry are weighted average treatment. The formula is as follows:

$$\overline{wage} = \sum_{i=1}^{3} \overline{wage_i} \bullet N_i \Big/ \sum_{i=1}^{3} N_i$$

6 Per Capita completed investment in Pollution Control in Industrial Sector (gov)

In this paper, the index of completed investment in industrial pollution control is adopted to represent the investment factor of government pollution control (GOV). These investments come from the fiscal expenditure of provinces and cities, reflecting the institutional power of local governments in pollution control.

7 Technology Market Turnover (rd)

Technology successful conversion is a complex process, and technology contract turnover is a reliable data source that is easy to quantify [11]. It is more appropriate to measure the technological development in the middle of China by the turnover of technology market.

As can be seen from the previous theory, CEV represents the carbon dioxide emission per unit value. The larger CEV is, the greater carbon dioxide emission per unit value is and the lower energy utilization rate is. The smaller the CEV, the smaller the carbon dioxide emission per unit value and the higher the energy efficiency of the region. In this paper, CEV indicators of six provinces in central China from 2004 to 2017 were selected to rank, and the current situation of emission reduction efficiency in central China was preliminatively understood.

TABLE II. CARBON DIOXIDE EMISSION

Rank	Province	*CEV*	*MAX*	*MIN*
11	Shanxi	712941	1030080.56	450160.38
2	Henan	461719.81	569382.98	299756.28
3	Anhui	267442.6	335116.65	156986.74
4	Hubei	230041.5	317168.17	161615.23
5	Hunan	210098.5	260988.76	121201.54
6	Jiangxi	128027.53	160455.95	79143.14

As can be seen from the above statistics, there is not only a big gap in the traditional endowment (capital stock of industrial sector) in central China, but also a provincial difference in government governance. Governments at all levels have a big gap in their investment in environmental pollution control and governance. The industrial production mode in the central region of China is still dominated by extensive production, with high output accompanied by high energy consumption and high pollution emissions, coupled with insufficient investment in environmental control and governance, there is still a long way to go to achieve "carbon peak and carbon neutrality" in the central region.

This part analyzes the impact of traditional elements, industrial internal structure, technological progress and government governance on unit value carbon dioxide emissions in central China from 2004 to 2017 by panel regression.

The panel model established in this paper is as follows:

$$lnCEV_t = lncapital_t + lnwage_t + lnrd_t + lnindus_t + lngov_t + \theta_t + \mu_t + \varepsilon_t$$

This empirical analysis adopts mixed regression, individual fixed effect model and random effect model.

TABLE III. OVERALL REGRESSION RESULTS IN CENTRAL CHINA

Dependent Variable CEV			
Model Form	*Model（1）*	*Model（2）*	*Model（3）*
lncapital	3.42* （2.10）	3.42*** （4.36）	-0.09 （-0.38）
lnwage	-2.27 （-1.06）	-2.27** （-2.27）	-0.23 （-0.87）
lnindus	-1.25** （-3.47）	-1.25*** （-3.05）	-0.11*** （-2.16）
lngov	1.60 （1.81）	1.60*** （3.74）	0.00 （0.03）
lnrd	-1.80** （-3.33）	-1.80*** （-5.61）	-0.06 （-0.65）
C	-0.25 （-0.02）	-0.2502834 （-0.03）	3.61 （2.58）

From model 2 in the above table, we can conclude that the regression result is very significant. On this basis, we conduct the analysis of the results :(1) the per capita capital stock of the industrial sector is positively correlated with CEV. CEV will increase by 3.42% when per capita capital stock of industrial sector increases by 1%. Therefore, the transformation of industrial internal structure is an important factor to reduce carbon emissions and improve carbon efficiency. (2) The per capita wage of the industrial sector is negatively correlated with CEV. When per capita wage of industrial sector increases by 1%, CEV will decrease by 2.27% on average, and labor cost's forcing effect on carbon emission efficiency can be fully reflected. (3) Industrial structure is negatively correlated with CEV. When the industrial structure increases by 1%, CEV will decrease by 1.25% on average. The optimization of industrial structure has a relatively obvious promoting effect on CEV. (4) There is a positive correlation between the per capita investment in pollution control and CEV. CEV will increase by 1.60% when the per capita pollution control investment of industrial sector increases by 1%. The per capita pollution control investment of industrial sector, as an indicator of government governance, has a lower impact on CEV than other factors. (5) As an indicator of technological progress, technological market turnover is negatively correlated with CEV. The empirical analysis of this index is consistent with the theoretical conclusion. There may be the following situations: First, there is a long cycle of technological progress, and there is accumulation and lag.

IV. CONCLUSIONS AND RECOMMENDATIONS

Based on the Marxist ecological theory, this paper introduces the political economy model and empirical analysis tools to investigate the factors influencing carbon dioxide emissions per unit value in central China and the differences in the contribution of traditional factor endowment, technological progress and industrial internal structure to carbon emissions per unit value.

Specific suggestions are as follows: First, formulate long-term governance management model, optimize the current environmental assessment indicators. In solving environmental governance problems, standards are often not unified, resulting in confusion. Secondly, the GDP-only assessment method also makes governments at all levels selectively ignore environmental problems and sacrifice the environment to preserve economic growth. Second, accelerate the transformation of industrial structure and promote the initiative of green industry. First of all, promote enterprise technological innovation, realize industrial structure upgrade, under the strong support and guidance of the government, actively cultivate green environmental protection, green energy saving and other related industries. Third, accelerate technological research and development, develop and popularize the use of new energy. First of all, increase investment in environmental research and development, establish international low-carbon technology transfer mechanism, the government should continue to strengthen support for environmental technology research and development, and actively formulate relevant preferential policies and incentives.

REFERENCES

[1] Shengyue Liu, Dynamic characteristics and influencing factors of carbon emissions from energy consumption in Guangxi. China Market,2021,{4}(16):7-12.]

[2] Krugman p. Myth of Asia's Miracle[J].The Foreign Affaires,1994,73(6):62-78

[3] Shi Pu. Why to Promote high-quality development in central China [N]. Learning Times,2021-08-06(001).

[4] Jie Wang, Zhiguo Li, Jijian Gu. Decoupling elasticity and driving factors of CARBON emissions and economic growth in BRICS countries: Based on Tapio decoupling and LMDI model [J]. World Regional Studies, 201,30(03):501-508.]

[5] Manzhi Liu, Xixi Zhang, Mengya Zhang, Yuqing Feng, Yingjie Liu, Jixin Wen, Liyuan Liu. Influencing factors of carbon emissions in transportation industry based on C[sbnd]D function and LMDI decomposition model: China as an Example [J]. Environmental Impact Assessment Review, 201,90:

[6] Xinyu Qi, Jinqiang Yan. Carbon emission constraints and economic growth: a theoretical and empirical study [J]. Academic monthly,2010,42(07):73-77.

[7] The Complete Works of Marx and Engels [M]. Volume 31, People's Publishing, 1976:124.

[8] Zhen Li, ZHANG Meng. Political economy analysis of China's regional low-carbon competitiveness: theory and demonstration [J]. Financial and economic research,2016,42(06):133-144.

[9] Jinhua Feng. Economic research journal,2012,47(01):31-41.

[10] Yibo Yang. Estimation of China's physical capital stock by industry (1980-2018)[J]. Shanghai Economic Research,2020(08):32-45.

[11] Baoji Du, Pengju Zhang. Research on the policy effect of scientific and technological achievements transformation: From the perspective of per capita turnover of technology market contract [J]. Journal of Liaoning university (philosophy and social sciences),2020,48(03):82-93+185.

(Note:Tingting Yin and Xuemei Long are the corresponding authors)

Digital economy has become a new engine for regional development

Zemin Hu, Jinghao Zhao*, Lu Shi
School of business
Guilin University of Electronic Technology
Guilin, People's Republic of China
e-mail: 492472199@qq.com

***Abstract*—In the context of the global epidemic and slowing economic growth, the digital economy, a new economic form based on the Internet and corresponding emerging technologies, has shown a vigorous development and has become a new engine for regional development. On the basis of combing the connotation and essence of digital economy, this article discusses the relationship between digital economy and regional development and the internal mechanism of digital economy to promote regional development.**

Keywords-digital economy; regional development; digitalization; technological innovation

I. INTRODUCTION

The concept of digital economy can be traced back to the 1990s. The American scholar Don Tapscott formally proposed the concept of digital economy in his book "Digital Economy". With the rapid development of Internet technology, the digital economy has been endowed with rich connotations.

The "White Paper on China's Digital Economy Development"[1] holds that the digital economy is an emerging economic form with digital knowledge and information as the key production factors, digital technology as the core driving force, and modern information network as an important carrier. Through the deep integration of digital technology and the real economy, the digital economy continuously improves the level of digitization, networking, and intelligence, and accelerates the reconstruction of economic development and governance models. It can be seen from this that the digital economy is a relatively broad concept. Any economic form that directly or indirectly uses data to guide resources and promote the development of productivity can be included in its category. At the technical level, the digital economy includes emerging technologies such as 5G, big data, artificial intelligence, the Internet of Things, cloud computing, and blockchain[2].

The essence of the digital economy is informatization[3]. Informatization is a social economic process in which the industrial economy turns to the information economy caused by the revolution of production tools such as computers and the Internet. Specifically, informatization includes the industrialization of information technology, the informatization of traditional industries, the informatization of infrastructure, and the informatization of lifestyles[4]. The core of digital economy is to change the original social production mode and business model by digitizing and informatization of all business elements, improve the operation efficiency of enterprise business activities, and make digital productivity form a harmonious relationship with traditional production relations as far as possible, so as to improve the comprehensive value of economic activities to human society as a whole.

II. THE RELATIONSHIP BETWEEN DIGITAL ECONOMY AND REGIONAL DEVELOPMENT

Regional development refers to a series of economic and social activities centered on resource development, industrial organization, and structural optimization within a certain time and space[5]. Regional development can be divided into three stages: primary stage, growth stage, and transformation stage. Regional development is embodied in economic, political, social, cultural and other aspects, and economic, political and social development is undoubtedly the main aspect. In the same way, digital economy, digital government, and digital society are the three main lines, three engines, and three major achievements of the digital age. The three promote and complement each other to form a flywheel of regional development and jointly promote regional development.

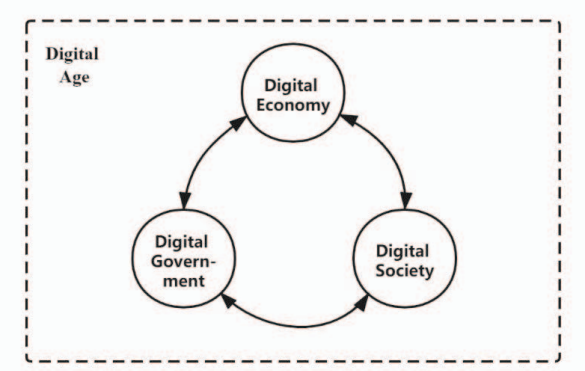


Figure 1. The three main lines of the digital age

The vigorous development of the digital economy has brought about a huge leap in productivity, has become the cornerstone of promoting regional digital development, and has provided material and technical conditions for digital government and digital society[6]. Digital government has improved the production relations of regional development and provided institutional guarantees for the development of digital economy and digital society; the construction of digital government will surely release more policies and reform dividends, and provide strong impetus for regional development in terms of systems and policies. The digital economy and digital government serve the digital society. The construction of the digital

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00012

society directly reflects the level of regional development. The development of the digital society will also significantly drive the development of the digital economy and force the construction of the digital government.

III. THE MECHANISM OF DIGITAL ECONOMY PROMOTING REGIONAL DEVELOPMEN

The development of the digital economy itself can promote regional development. The fast-growing digital economy promotes the continuous optimization of traditional economic models, accelerates the reshaping of global business models, and drives the intensive integration, collaborative development, efficient utilization, and network sharing of capital, labor, technology and other production factors[7-8]. The penetration of the digital economy into traditional industries can further promote economic development. In the digital economy, the production costs of enterprises have strong high fixed costs and low marginal costs. The traditional production costs of enterprises include high fixed costs and relatively high marginal costs. Marginal costs gradually decrease at the beginning, but after reaching a certain scale, diseconomies of scale begin to appear, and marginal costs begin to rise. But for the digital economy, its marginal cost gradually decreases to zero as the scale expands.

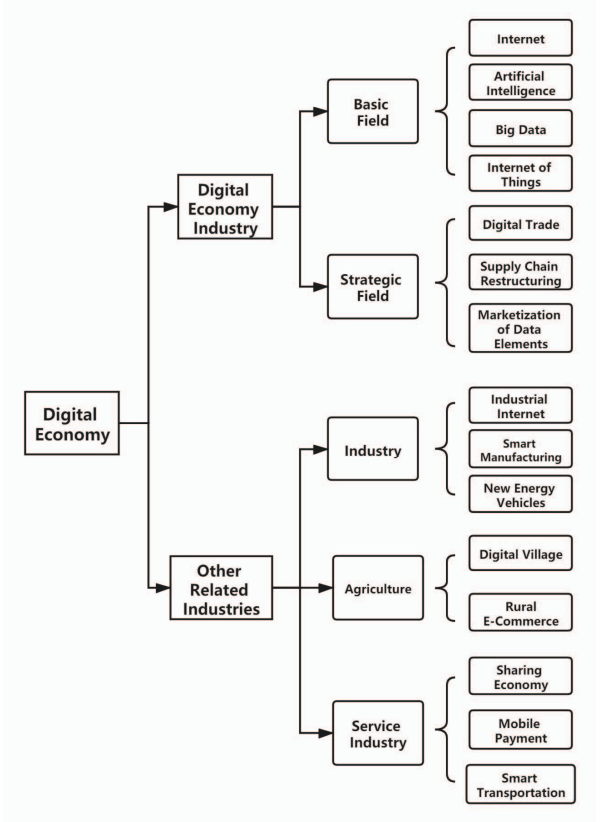


Figure 2. The basic framework for digital economy to promote regional development

On the basis of the continuous innovation of smart technology and smart devices, the industrial Internet is developing rapidly with the Internet of Things, cloud computing and big data as technical means. This can not only significantly improve the efficiency of resource utilization, but also promote the high integration of all links in the manufacturing industry chain, form a new data realization model, and promote the transformation and upgrading of the industrial structure. At the same time, the essence of the development of the digital economy is to increase the flow of information and knowledge elements in the entire economic system. In this way, it promotes the refinement of the regional spatial division of labor and the improvement of the efficiency of inter-regional transactions, optimizes the structure of the division of labor between regions, and realizes the transformation and upgrading of the inter-regional economic structure. In addition, the development of the digital economy can significantly increase the efficiency and depth of transactions between enterprises and customers, continuously improve the supply and demand structure of the market, and promote the transformation of the economic structure.

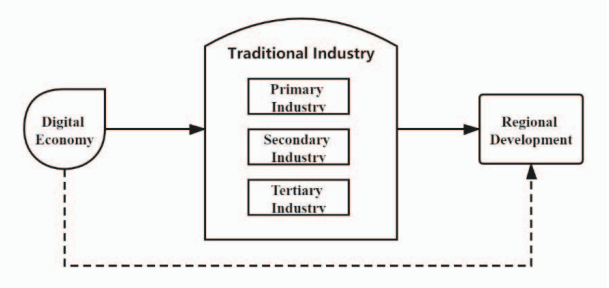


Figure 3. Digital economy, traditional industries and regional development

The development of the digital economy can also promote the integration of digital basic industries and other industries[9]. Basic industry of the digital economy itself is a key component of the national economy. The development of the basic industry of the digital economy has driven the improvement of the industrial structure. Boundaries between the three traditional industries have gradually disappeared, and the development of industries has shown a trend of cross-convergence[10], from the traditional labor-intensive, capital-intensive, and technology-intensive to digital-intensive transformation, which promotes the transformation and upgrading of traditional industries and significantly improves production efficiency. This will eventually lead to the development of the region.

IV. CONCLUSIONS AND RECOMMENDATIONS

In the new round of global industrial and technological revolution, the integration of the digital economy, represented by the Internet, artificial intelligence, blockchain, cloud computing, and big data, with many social and economic fields has become an irresistible trend of the times. This article analyzes the relationship between the digital economy and regional development and the internal mechanism that promotes regional development. And this article points out that the development of digital economy itself can promote regional development and that the digital economy can also promote regional development by driving the integration of digital basic industries and other industries.

In order to play the role of the engine of the digital economy, it can be implemented from two aspects. One is to vigorously develop industries related to the digital

economy. The government must increase support for the digital economy and give full play to its due guidance and support functions; enterprises must increase R&D and investment in the digital economy, and increase the output of innovation results. The second is to actively promote the digital transformation of traditional industries. Digital technology can be used to upgrade traditional industries, and a number of high-level digital transformation demonstration projects can be built in the region to deepen the digital application of manufacturing, operation management, and market services.

REFERENCES

[1] China Academy of Information and Communications Technology. White Paper on China's Digital Economy Development [M]. Beijing: China Academy of Information and Communications Technology,2020:3.

[2] Ruiyun Du, Baoyi Wang. New retail: research review and prospects [J]. Business Economy,2020,39(08):128-135.

[3] Delin Sun, Xiaoling Wang. The nature and late-comer advantages of digital economy [J]. Contemporary Finance and Economics,2004(12):22-23.

[4] Xiaozhou Hu. Research on Shenzhen Hi-tech Industrialization [D]. Southwestern University of Finance and Economics,2002.

[5] Lin Zhao. Research on Risk Evaluation and Control of Real Estate Development Projects Based on the Whole Life Cycle [D]. Harbin Engineering University,2012.

[6] Xue Li, Fuxiang Wu, Lile Zhu. Digital Economy and Regional Innovation Performance [J]. Journal of Shanxi University of Finance and Economics,2021,43(05):17-30.

[7] Juan Wang. Digital economy drives high-quality economic development: factor allocation and strategic choices [J]. Ningxia Social Sciences,2019(05):88-94.

[8] Jiangbo Han. Innovation drives high-quality economic development: the mechanism of factor allocation and strategic choices [J]. Contemporary Economic Management,2019,41(08):6-14.

[9] Zhaojie Liu. Research on Industrial Convergence in the Context of Digital Economy [D]. University of International Business and Economics,2018.

[10] Xiaoyin He, Tianting Cheng. The strategic choice of digital economy to promote high-quality economic development [J]. Business Economics Research,2021(10):189-192.

Analysis and Decision-making of Regional Economic Vitality

Chunan Zhang[1], Ying Gan[2], Hao Yang[3], Bingyun Wang*, PingShao*
School of Atmospheric Science
Chengdu University of Information Technology, Chengdu 610225, China
*Correspondence Author: Bingyun Wang，E-mail：bywang@cuit.edu.cn

***Abstract*—In order to study the impact of different leaders on the U.S. economy, this paper uses a gray correlation model to obtain the more sensitive and prominent factors among all the factors. After that, these factors are used to calculate the corresponding multiple linear regression equations to predict the U.S. economy.**

Keywords-advanced correlation analysis, multiple linear regression, gray prediction model

I. INTRODUCTION

The U.S. presidential elections are held every four years. 2020 is the annual U.S. presidential election, with Republican candidate Donald Trump and his opponent, Democrat Joe Biden, running for president. The candidates of the two parties have different political positions on key development areas such as finance and trade. The election of the different candidates will have a large impact on the U.S. economy as well as the global economy.

Economic dynamics is an important indicator that can reflect the economic situation and economic development potential, and the gray model can be better applied to predict the development trend of GDP in the future[1].Analyzing the impact of different indicators [2] on the U.S. economy through the gray correlation method and used the principal component analysis to extract the main factors[3],which finally gave the quantitative calculation of economic vitality indicators.

Therefore, it is important to objectively and compreh - ensively evaluate the impact of the policies of the two candidates on the economy and explore the reasons for the differences, in order to have a greater impact on the U.S. economy as well as the global economy. In this paper, our study intend to analyze and model the impact of different factors on economic development and use them to make predictions for future U.S. economic development

II. MODEL BUILDING

.Gray correlation analysis

1)Determine the reference series and comparison series.

The data series reflecting the characteristics of the system behavior is called the reference series. The data series consisting of factors that affect the system behavior is called the comparison series.

2)Dimensionless processing of the reference series and comparison series

The physical meaning of the factors in the system is different, resulting in the data not necessarily having the same magnitude, which does not facilitate comparison.

3)Finding the gray correlation coefficient of the reference series and the comparison series

The degree of correlation is, in essence, the degree of difference in geometry between the curves. Therefore, the size of the difference between the curves can be used as a measure of the degree of correlation. For a reference series X_0, there are several comparative series $X_1, X_2, \ldots, X_n$ .The correlation number $\xi(X_i)$ between each comparative series and the reference series at each moment (i.e., each point in the curve) can be calculated by the following formula: ρ is the coefficient of discrimination, usually between 0 and 1, usually 0.5

$\Delta\max$ is the maximum difference between the two levels.

To absolute difference between each point on the Xi curve of each comparison series and each point on the X0 curve of the reference series, and is given as $\Delta_{oi}(k)$.

So the number of contacts $\xi(X_i)$ can also be simplified as the following equation：

$$\xi_{oi} = \frac{\Delta(\min) + \rho\Delta(\max)}{\Delta_{oi}(k) + \rho\Delta(\max)}$$

4)Find the degree of relevance

Because of the correlation series correlation:

$$r_i = \frac{1}{N}\sum_{k=1}^{N} \xi_{oi}(k)$$

r_i - compares the gray correlation of a series X_i to a reference series X_0, or serial, average, or line correlation.

The closer the r_i value is to 1, the better the correlation.

5)Relevance

The degree of association between factors is mainly described by the order of the degree of association, not just the size of the association. The correlation degree of m subsequences to the same parent sequence is arranged in the

Co-first-authors：Chunan Zhang, Ying Gan, Hao Yang.

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00013

order of size, then the correlation sequence is formed as $\{x\}$, which reflects the "superiority" relationship of each subsequence to the parent sequence.

If $r_{0i} > r_{0j}$, then $\{x_i\}$ is said to be better than $\{x_j\}$ for the same parent sequence $\{x_0\}$, and is recorded as $\{x_i\} > \{x_j\}$; r_{0i} represents the eigenvalue of the subsequence.

Gray correlation analysis [4] is to consider the factor values of the object of study and the influencing factors as points on a line, compare them with the curve drawn by the factor values of the object to be identified and the influencing factors, compare the closeness between them, and quantify them separately to calculate the correlation degree of closeness between the object of study and the influencing factors of the object to be identified, and judge the degree of influence of the object to be identified on the object of study by comparing the magnitude of each correlation degree.

.Multiple Linear Regression

1)Create multiple linear regression equations

Due to the multifaceted nature of the relationship between things, changes in a dependent variable may be influenced by multiple other independent variables, so multiple linear regression equations are created.

2) Basic Model and Steps：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + e$$

The above equation indicates that the variable y in the data can be approximated as the linear function of independent variable $x_1, x_2, \ldots, x_m$.

β_0 is a constant term, β_m is the partial regression coefficient，which represents the average change in when x_j increases or decreases by one unit, holding the other independent variables constant, e is the random error (residual) after removing the effect of m independent variables on y.

3) General steps

a.Find the partial regression coefficient$b_0, b_1, b_2, \ldots\ b_m$

b. Derive the regression equation：

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m$$

c. Test and evaluate the regression equation and the effect of each variable.[5]。

.GM(1,1)(Grey prediction model)

1)The basic principle

Gray Prediction identifies the degree of disparity in development trends between system factors, that is, performs correlation analysis, and generates and processes the original data to find the rules of system changes, generates data sequences with strong regularity, and then establishes corresponding differential equations to predict the future developments. It constructs a Gray Prediction Model with a series of quantitative values of the characteristics of the prediction object observed at equal time intervals and predicts the feature quantity at a certain time in the future, as well as the time it takes to reach a certain feature quantity.

2)Basic steps

a. Data inspection and processing

To ensure the feasibility of the $GM(1,1)$ modeling method, it is necessary to check the known data

.Set the original data column as $x^{(0)} = (x^0(1), x^0(2), \ldots, x^0(n))$，calculate the ratio of the sequence：

$$\lambda(k) = \frac{x^{(0)}(k-1)}{x^{(0)}(k)}$$

If all the grade ratios fall within the acceptable coverage interval, the sequence $x^{(0)}$can establish a $GM(1,1)$ model and can perform gray prediction. Otherwise, perform appropriate transformation processing on the data, such as translation transformation:

$$y^{(0)}(k) = x^{(0)}(k) + c \quad , k = 1,2, \ldots, n$$

c is taken so that the rank ratio of the data column falls within the acceptable coverage.

b. Establish $GM(1,1)$ model

May wish to set $x^{(0)} = (x^0(1), x^0(2), \ldots, x^0(n))$ to meet the above requirements, and use it as a data column to establish a $GM(1,1)$model.：

$$x^{(0)}(k) + a x^{(1)}(k) = b$$

Use regression analysis to obtain the estimated values of a and b, so the corresponding whitening model is :

$$\frac{dx^{(0)}(t)}{dt} + a x^{(1)}(t) = b$$

Solution to：

$$x^{(1)}(t) = \left(x^{(0)}(k) - \frac{b}{a}\right) e^{-a(t-1)} + \frac{b}{a}$$

So get the predicted value:

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{b}{a}\right) e^{-ak} + \frac{b}{a}$$

To get the predicted value accordingly:

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k)$$

.

III. MODELS SOLUTION

I. Models Solution

(1)Grey correlation method to extract key factors.

Based on data provided by the World Bank in various parts of the United States from 2008 to 2018, 8 factors have been selected initially that have a greater impact on US GDP (the proportion of fossil fuel energy consumption, railways (total kilometers), renewable energy power generation and energy consumption). Screening these 8 factors by gray correlation degree with MATLAB software, and then their correlation degrees are shown in Table 1.

Factors	Employment rates	railway	Renewable power generation	Science and technology research
Correlation coeffient	0.6946	0.4681	0.5193	0.6776
Factors	taxation	Commodity trade	Fossil fuels	Energy consumption
Correlation coeffient	0.7272	0.9368	0.7115	0.6522

Table 1. The gray correlation degree

From this figure, a 5-factor correlation filtered from these 8 factors can form obviously. It includes commodity trade, taxation, fossil fuels, employment rates, science and technology research and development expenditures, all of which are associated with a greater likelihood of economic growth.

(2)Gray prediction for five-factor

Using the gray correlation method, the five factors with a greater degree of correlation applied to gray prediction models later can be screened out. The specific prediction result is as follows:

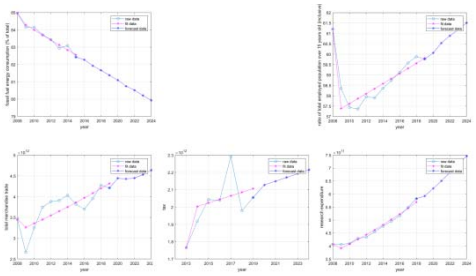


Figure 1. The five factors are respectively used to predict the results of the gray prediction model

To predict their impact on future GDP, the forecast data obtained from each factor of our forecasts are substituted for Trump and Biden into the multiple regression equation (Below are the predictors of the two different leaders: x1, x2, x3, x4, and x5 represent the proportion of fossil fuel energy consumption, commodity trade, technology research and development expenditure, employment rate, and taxation) Based on 19 years of data, the following rationalized predictions based on Biden's policies during the general election have been made: x1 (fossil fuel) decreases by 3% each year, x2 (trade volume) increases by 3% each year, x3 (research and research expenditures) An annual increase of 3%, x4 (the employment rate) an annual increase of 0.635% (based on the population of the United States), and x5 (tax revenue) an annual increase of 3%.

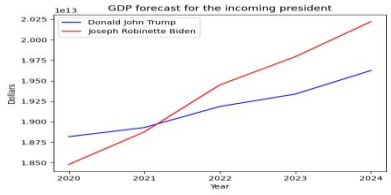


Figure 2. GDP forecast for the incoming president

II. Model Adaptability Test

An improved neural network model for sensitivity analysis. Figure IV below shows the average relative residuals and rank ratio bias tests. The results show that the five factors fit well.

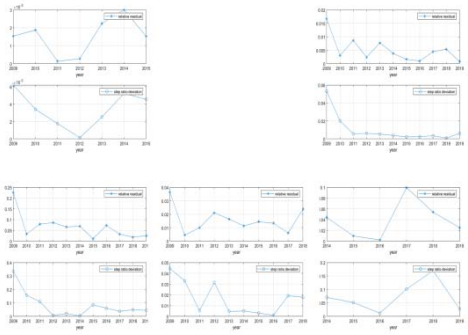


Figure 3. Test of average relative residual and grade deviation

IV. .CONCLUSION

.In summary, in the long run, Biden's policies are good for the U.S. economy. However, in the short term, Biden's policies have not worked as well as Trump's policies. In the long run, Biden's policies are far better than Trump's policies in terms of the long-term development of the United States. His presidency will benefit the U.S. economy. It is not difficult to conclude that new energy sources and energy conservation and environmental protection have huge economic benefits and prospects, although they do not bring the same benefits as traditional energy sources in the short term

ACKNOWLEDGMENT

A Project Supported by Chengdu University of

Information Technology (CUITGOMP202102), and Innovative training program for college students (Code: CUIT202110621023).

REFERENCES

[1] *Zhao Ying. Study on the development of rural tourism and rural culture coupling relationship in Hainan in the promotion of new urbanization[J]. Modern Business,2019(07):80-82. .*

[2] *Su Jing,Zhi Yihao. Analysis of regional economy and house prices based on gray correlation method[J]. Journal of Science and Technology Economy,2021,29(08):25-26..*

[3] *Zhu Shilin, Liu Yang, Guo Zhuoying, Guo Zefeng. Analysis of regional economic dynamics and its influencing factors using PSO method--in the context of the 2019 Asia-Pacific University Student Mathematical Modeling Competition B-title[J].Journal of Guangzhou Maritime Academy,2020,28(04):73-78.*

[4] *Liang Wei, Wang Duanxia, Zhang Li. Gray correlation analysis method. WanFang, 2013.*

[5] *Ye Zhonghua. Application of gray prediction model in mathematical modeling [J]. Journal of Science and Technology Economy,2018(33):199-199.*

[6] *Chen Hui,Hao Xinyu. Research on regional economic vitality based on multiple regression analysis[J]. China collective economy,2020(30):7-8.*

Research and Application of Community Pension Service Data Analysis System Based on Data Mining Algorithm

Xiaofeng Yao[1], Cheng Jiang[1]

（1. Jiangsu Key Lab of IoT Application Technology, Wuxi Taihu University, Wuxi 214064，China）

Abstract: In recent years, with the development of China's aging population, pension issue has become a major issue related to the national economy and people's livelihood continues to be concerned. This paper aims to study the design and implementation of data analysis system of community elderly care service based on data mining algorithm. Firstly, this paper introduces the theory of data mining, and designs three modules of community pension service. Secondly, the gender, age, education level, the number of people living with the elderly and the situation of chronic diseases are used as variables for model design, and finally a community service pension system is designed in accordance with the national conditions. The experimental results of this paper show that, on the basis of data mining algorithm, the construction of community pension service system satisfies the needs of the elderly, among which, the highest reaches 91.8%, which also shows that the pension service system brings great convenience to the elderly.

Keywords: *Data mining algorithms, community pension services, data analysis systems*

I. INTRODUCTION

With the development of the information age, data mining applications are very extensive, including a variety of technologies, of which classification technology is currently one of the most critical technologies in data mining [1]. Classification is mainly to establish a classification model based on sample data sets to predict the development trend of a large number of data sources. Due to the importance of classification technology, it has attracted widespread attention from the scientific and industrial circles [2-3]. This paper mainly implements an intelligent data analysis platform that analyzes and processes data under the big data platform, cluster deployment and management, and data migration, thereby assisting decision-making [4]. The realization of data mining algorithms under the big data platform is of great significance for realizing business model-based big data construction and big data service applications [5].

With the development of society and the advancement of technology, more and more researchers have devoted themselves to the research of data mining algorithms [6-7]. Among them, it usually takes a long time to build a machine learning model, but none of the existing machine learning software can provide an important progress indicator [8]. Similarly, running data mining algorithms usually takes a long time, but none of the existing data mining software provides important progress indicators. Gang L considers the problem of providing progress indicators for the construction of machine learning models and the execution of data mining algorithms [9]. He discussed the inherent goals and challenges of this issue. He then described the initial framework for implementing such progress indicators and their two high-level potential uses, with the goal of stimulating future research on the subject [10]. His research direction is very in line with current needs, but it is still obviously insufficient in terms of practicality.

The innovation of this paper is (1) analyzes the current aging society. With the intensification of aging, the growth rate of the elderly and the empty-nest elderly are also accelerating, and the necessity of building a smart community elderly care service system is analyzed. (2) When conducting statistical analysis on the demand for community elderly care services, cross-link the influencing factors and the demand for various specific services to analyze the difference in demand.

II. COMMUNITY ELDERLY CARE SERVICE METHOD BASED ON DATA MINING ALGORITHM

A. Data Mining Algorithm

Data mining, also known as knowledge discovery, refers to intelligently searching for valuable information hidden in large amounts of imperfect, noisy, fuzzy and random data.

The role of data mining is to discover or extract potentially useful information from a large amount of data. The development process can be divided into five stages. The first stage is based on an independent system, using vector data, and only supports one algorithm. The second stage is to combine data mining with databases and support multiple algorithms at the same time. The third stage of data mining is based on a grid-based calculation method. At the same time, the prediction model is integrated into the data mining process, which can process Web data and semi-structured data. The fourth stage is distributed data mining, which is a way of

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00014

distributing multiple algorithms for execution on multiple nodes. The fifth stage is data mining based on cloud computing, which uses a distributed parallel processing and service model. The same algorithm can be distributed on multiple nodes, and multiple algorithms can also be executed in parallel. The computing resource structure of multiple nodes will be centralized to form a shared resource pool, and tasks will be allocated on demand.

B. Data Mining K-means Algorithm

The K-means algorithm is a partition-based algorithm, and it is also one of the most widely used algorithms in clustering algorithms. Its basic idea is: give the value of K in advance, and then randomly select K points from the original data set as the initial cluster center. Then calculate the distance between it and all cluster centers for each point in the data set, and assign it to the cluster represented by the nearest cluster center. Finally, when all the points are divided, the center of the class is updated and a new round of iterative operation is performed. Repeat the iteration process as described above until the cluster centers converge or meet other conditions (usually the maximum number of iterations).

In the description of the K-means algorithm flow, when all the points in the data set are assigned, the center of the cluster is updated, and the new cluster center is finally obtained as:

$$c_j^* = \frac{1}{N_j}\sum_{u_m \in c_j} u_m, 1 \le j \le K \quad (1)$$

The process of assigning points and updating cluster centers is repeated until the objective function converges, that is, the center point no longer changes, or the maximum number of iterations is reached.

C. Community Elderly Care Services

The so-called community elderly care refers to "the government and social forces rely on the local society to provide living care, housekeeping services, rehabilitation care, and mentally comfortable service forms for the elderly at home." The content should include lifelong care services, medical services, emotional services, cultural and entertainment services, and services for the elderly. Communities establish day care centers for the elderly, set up community health service centers, and open elderly activity centers to provide elderly care services.

The community elderly care service system conforms to my country's national conditions and is an extension of other elderly care models within the framework of the social system. The safety of the elderly is the foundation, and secondly, it is necessary to take the age management of the community as a core, based on the institutions that provide various age services, and take care of the elderly at home as the goal. It is provided by volunteers. The basic value of free services is for the elderly in the whole society, so that the elderly can enjoy a full range of elderly care services, protect the personal safety and health of the elderly, and enrich the lives of the elderly on the other hand. The system design is mainly divided into three modules, namely information management module, elderly service module, community announcement and health express module. The system structure diagram is shown in Figure 1:

Figure 1. Community elderly care service system framework

III. EXPERIMENTAL DESIGN OF COMMUNITY ELDERLY SERVICE

A. Model Design

In the following, five factors including gender, age, education level, number of people living together, and prevalence of chronic diseases are used as independent variables. Assign values to variables and establish a Logistic regression model to analyze the factors affecting the demand for community care services.

The logistic regression model is:

$$\text{logit}(n) = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n + u \quad (2)$$

Among them, n is the probability of the dependent variable, $x1$、$x2$ …… xp is the explanatory variable, b is the coefficient, and u is the random disturbance term. Gender, age, educational level, number of people living with chronic diseases are assigned values according to the data situation and research needs, as independent variables, and the needs of life care services, medical care services, and leisure and social services are separately assigned.

B. Construction of An Evaluation Index System for the Quality of Community Elderly Care Services

Community services for the elderly are divided into four types of resource services: daily care, medical care, rehabilitation, and spiritual culture. Each resource has its own characteristics, and the quality evaluation index of each resource is also different. For example, daily services focus on concentration, entertainment, and other pleasures. Daily core life indicators, medical services pay more attention to medical equipment, doctor qualifications, etc. In view of the different service resources of the elderly community, it is

necessary to establish the corresponding quality of the elderly community service.

The construction of the community elderly care service quality index system is divided into three steps: the selection of evaluation indexes, the determination of index assessment standards, and the calculation of index weights. First of all, according to the objective, independent, scientific, and comprehensive indicator selection principle, through questionnaire surveys, the needs of the elderly in social life are taken as the starting point, as the service elements of the elderly service providers; the service process and the elderly services is the starting point, as well as the feelings and expectations of the elderly about the service. Secondly, the quality evaluation indicators of community care services, rules and regulations, and relevant laws and regulations, clauses and documents of the indicator content are used to determine the standards of the evaluation indicators; and according to the analytic hierarchy process, the indicators for evaluating various resources of community nursing services, calculating matrices, and calculating The average feature weight of each index in the evaluation table, check the consistency of the evaluation table, and constantly change the weight value of each index according to the consistency result, until the weight meets the requirements.

IV. COMMUNITY ELDERLY CARE SERVICE SYSTEM BASED ON DATA MINING ALGORITHM

A. Current Status of Community Elderly Care Services

Community elderly care services have developed against the background of the social elderly dependency ratio increasing year by year and the introduction of relevant policies. In recent years, community service agencies have increased year by year and community elderly care services have been developed, providing favorable conditions for the improvement of community elderly care services.

According to statistics released by the National Bureau of Statistics in 2019, from 2010 to 2019, the number of people over 65 increased from 100.24 million to 138.24 million, an increase of 38 million in 10 years. The ratio of the population aged 65 or older to the total population increased from 7.52 percent in 2010 to 9.83 percent in 2019, and the ratio of elderly dependants increased from 10.32 percent in 2019 to 13.75 percent in 2019.

The old-age dependency ratio is "an important indicator for judging the economic burden of the elderly population on society and measuring the economic burden of laborers. It reflects the impact of the elderly population on social and economic development. The elderly dependency ratio in 2019 is shown in Figure 2:

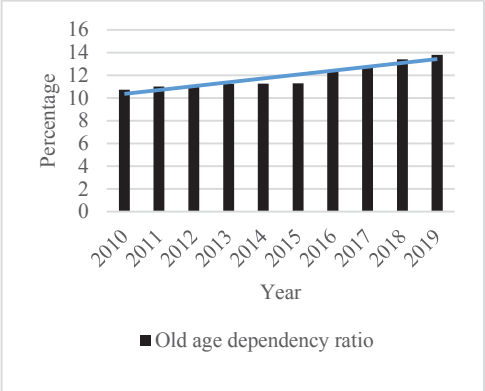


Figure 2. Proportion of elderly support from 2010 to 2019

From the data in the figure, it can be seen that from 2010 to 2019, my country's old-age dependency ratio has increased year by year, and the old-age dependency ratio has increased year by year, indicating that the society's support burden on the elderly has increased. It shows that my country has already entered an aging society, and the degree of aging is deepening. The deepening of the aging degree will put pressure on my country's economic operation and social stability. The improvement of elderly care services is imminent. Community elderly care services are an important part of elderly care services. It provides the elderly in the community with pension insurance including life care, medical care, and leisure and social life. When improving community elderly care services, in addition to relying on national power, social forces can also be fully mobilized to participate in the construction and provision of community elderly care services, relieve social pension pressure.

B. Overall Demand for Community Elderly Care Services

Community elderly care services in this article include life care services, medical care services and leisure social services. The overall demand analysis of community elderly care services is the frequency statistics and comparison of the overall demand for these three types of services.

Table 1. Demand for community elderly care services

Service type	Frequency	Proportion (%)
Life care	435	77.1
Medical care	260	46.1
Casual social	518	91.8

It can be seen from Table 1 that among the 564 interviewees, leisure and social services have

the largest number of choices and the highest demand. There are 518 people in need, reaching 91.8%, followed by life care services, reaching 77.1%, and lastly medical care services. Accounting for 46.1%, shows that the elderly pay more attention to the satisfaction of leisure life needs, while participating in leisure social activities, improve life satisfaction and meet social needs.

V. CONCLUSIONS

On the basis of survey data, this paper selected the 564 research samples, analyses the sample characteristics, building form statistics to calculate the community endowment service demand, and community endowment service can be divided into life care services, health care and leisure social services three categories, through data mining algorithm of three types of service demand influence factors were analyzed. The construction and improvement of community pension service is an effective and necessary measure to deal with the aging of population. The analysis and research on the demand for community pension service provides a basis for providing targeted community pension service, which is conducive to improving the quality and efficiency of community pension service.

REFERENCES

[1] Daniel, Lane, Shima, et al. Adaptation Decision Support: An Application of System Dynamics Modeling in Coastal Communities[J]. International Journal of Disaster Risk Science, 2017, 8(04):374-389.

[2] Nichols T E , Das S , Eickhoff S B , et al. Best Practices in Data Analysis and Sharing in Neuroimaging using MRI[J]. Nature Neuroscience, 2017, 20(3):299-303.

[3] Li Z , Li H , Feng Y . Research on big data mining based on improved parallel collaborative filtering algorithm[J]. Cluster Computing, 2019, 22(5):1-10.

[4] Cui Z , An F , Zhang W . Internet Financial Risk Assessment Based on Web Embedded System and Data Mining Algorithm[J]. Microprocessors and Microsystems, 2021, 82(3):103898.

[5] Jia X , Nan Z , Han B , et al. Assessment on personal exposure to particulate compounds using an empirical exposure model in an elderly community in Tianjin, China[J]. Science of The Total Environment, 2016, 572(dec.1):1080-1091.

[6] Ling G , Xue C , Min Z . The Study on Community Health Education of Empty Nest Elderly[J]. Engineering, 2016, 05(10):137-139.

[7] Wang, Dan, Zhang, et al. Qualitative analyses of lived experience for residents in the Elderly Care Departments at the community health service centres in Southwestern China[J]. Health & social care in the community, 2018, 26(1):e164-e172.

[8] Mi H , Huang X , Anushya M , et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements[J]. Nucleic Acids Research, 2017(D1):D183-D189.

[9] Gang L . Toward a Progress Indicator for Machine Learning Model Building and Data Mining Algorithm Execution: A Position Paper[J]. Acm Sigkdd Explorations Newsletter, 2017, 19(2):13-24.

[10] Girardin, Jean-Louis, Hans, et al. The Actigraph Data Analysis Software: I. A Novel Approach to Scoring and Interpreting Sleep-Wake Activity[J]. Perceptual and Motor Skills, 2016, 85(1):207-216.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Machine Learning Techniques to Predict Academic Performance of Health Sciences Students

Hana Alharthi, PhD.
Dept. of Health Information Management & Technology. College of Public Health
Imam Abdulrahman Bin Faisal University (IAU)
Dammam, Saudi Arabia
e-mail: Halharthi@iau.edu.sa

***Abstract*—Prediction of academic performance of health sciences students prior to being fully engaged in academic studies will identify those students who may need early intervention. Machine learning (ML), a branch of artificial intelligence, can be used to predict the academic performance of such students and the factors that continue to impact their academic performance. *Objective*: To use a best fit model in ML to predict the academic performance of health science students and rank the most important factors affecting their performance. *Method*: The academic records of 3468 students were extracted from the student information system (SIS), which included preparatory year great point average (GPA), high school GPA, Achievement Test (AT), General Aptitude Test (GAT), and cumulative GPA upon graduation. Multiple machine learning algorithms were used to develop the best fit model to predict students' performance GPA and identify factors that contributed to GPA. *Results*: The best performing classifier based on area under the curve (AUC) is random forest (.773) followed by naïve bayes (.758), Support Vector Machine (.686), k-nearest neighbors (.684) and decision tree (.658), the three scoring methods showed preparatory year GPA, gender, and high school GPA were the top variables predicating student cumulative GPAs. *Conclusion*: Random forest model can assist college administrators and faculty in health colleges to predict which students are more likely to underperform during their undergraduate studies.**

Keywords: Machine learning; algorithms; Classifiers; GPA ; ML.

I. INTRODUCTION

Anxiety profoundly affects quality of life of those affected by it. Universities continue to invest in resources to support the academic performance of their students. Yet the students that may need it the most are less likely to seek this support. As such, it is helpful to identify students that may struggle and underperform, especially as it relates to students in the biomedical fields who require strong academic performance to navigate the next step of their career whether further training or entering the workforce. Harnessing information technology can be a powerful tool to identify such students. Institutions have a breadth of student databases that can be easily saved, extracted, examined, and analyzed. Educational data mining (EDM) is a sub-field of data mining that recognize patterns that are not known before by utilizing smart algorithms [1]. As a result, educators can gain insights into their own academic environment that leads to better understanding on how to improve student academic performance, decrease number of failing students and decrease percentage of dropouts [2].

Data mining is utilized in many sectors such as in economics, business, health, retail, to name a few. The education sector has also used data mining [3]. To predict students' performance or retention rates among other issues. For example, current cumulative GPA for engineering students was used to predict their final semester GPA [4]. Another study used data mining to predict students at risk of dropping out [5].

Although numerous studies have targeted failing students to reveal correlates of failing and/or predicting which student will fail, we aimed to focus on students who are not failing yet are low performers, those with a "C" or " D" GPA, to provide educators with helpful information to establish the needed resources to maximize the performance of such students. Specifically, we evaluated student preparatory year GPA (first year), high school GPA , Achievement Test (AT) and the General Aptitude Test (GAT) and demographics to predict those students who are likely to graduate with a 'C" or "D" GPA from a health field (medicine, dentistry, nursing, applied medical sciences, pharmacy, and public health). We employed the orange data mining software to develop the best fit model to predict student GPA upon graduation.

II. LITERATURE REVIEW

Although admissions to health sciences colleges is highly competitive and is based on high standards of previous academic performance (e.g., high school GPA; high scores in standardized tests, etc.), some students struggle, underperform, and graduate with lower cumulative GPA. Therefore, capturing these low performers at early stage and providing them with appropriate support may make a difference to elevate their academic performance and aid in their future success, whether to continue to graduate school in some cases or enter their respective career sector.

Measuring student performance would enable decision makers and stockholders (management, administrative stuff, and faculty) to have the right information to enable them to make the best decisions for their students. This would rely on solid data that can predict student performance and reliance

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00015

on data mining. Numerous studies have addressed this issue, some focused on overall performance over long period of time (>1 year) while others evaluated shorter periods (one course or one semester) [6]. Course participation and perceptions examined for 533 students in their first year of study, classified them into three categories of academic performance , low-risk, medium-risk, and high-risk who are more likely to fail or drop out [7].

Multiple research work has used cumulative GPA as the main predictive attribute to assess student performance and most found that high school GPA, gender, ethnicity, quantitative SAT scores, verbal SAT scores, significantly impacted students' graduation [8],[9]. Further age, gender, parent's marital status, parent's qualification, cumulative GPA of 7500 students was evaluated to predict students' academic performance in their first year of college [10]. Several studies predicted students' final GPA using students grades information [11],[12]. Educational data mining (EDM) is a sub- branch of data mining that is based on statistics and machine learning to explore and investigate education data. It is applied in different aspects of the branch of education such as academic performance, admission, graduation, retention, gifted students among others [13].

EDM is utilized mostly to measure student performance [14]. Further, machine learning algorithms are used to analyze education data. These algorithms would provide models to identify students at risk and provide early warning alerts for the university staff to take the necessary actions. For example, decision trees (DT), neural networks (NN), and support vector machine (SVM) produced models that predicted student dropouts [15]. Another study that predicted student's dropout based on cumulative GPA was modeled using k-nearest neighbors (KNN) with 87% accuracy [16]. To predict students time of graduation naïve bayes algorithm was used which produced a model with 70.83% accuracy [17]. Decision tree classifier predicted student success or drop out with 60.5% accuracy using socio-demographic variables in addition to major and courses being studied [18]. The same algorithm was used to predict student first year performance in a business informatics by using their high school state exam marks and first year success , the model produced accuracy of 76.65% [19] .

Another study used neural network algorithm to predict student performance in a specific course and was 92.3% accurate [20]. Support vector regression algorithm outperformed multiple linear regression using socioeconomic and university academic information to predict student' academic performance [21].

III. METHODOLOGY

A. Data collection

Imam Abdulrahman Bin Faisal University, in Dammam Saudi Arabia (IAU) has six health colleges: College of Medicine, College of Dentistry, College of Nursing, College of Applied Medical Sciences, College of Public Health, College of Applied Medical Sciences – Jubail. Admission requires high school students to take two standardized tests. One is the Achievement Test (AT) which test their knowledge in Biology, Chemistry, Physics and Math. The other is the General Aptitude Test (GAT) which measures analytical and deductive skills. The equation that determines their admission is 30% high school GPA, 40% AT, and 30% GAT. We collected data from 3468 student records across these colleges using institutional student information system (SIS) from 2012-2019 and data placed in excel sheet. The dataset has seven features and one target variable with four values as shown in table 1.

B. Data preprocessing

One column in excel sheet was created to further formulate the research question. A formula was used to create grade codes for the cumulative GPA based on university policy as shown in table 2. The grade code becomes the class variable with values "A", "B" ,"C" or "D". If class values is "C" or "D" then student need counselling otherwise, no need for counselling.

Table 1. Students related variables

Attributes	Description
Gender	M, F
PREP_GPA	GPA after first year of study
High School_ GPA	GPA upon high school graduation
GAT	General Aptitude Test
AT	Achievement Test
College	A health college
Year	Admission year
Grade Code (target)	Class (A,B,C,D)

Table 2. Grade point average (GPA) scale

Code	GPA out of 5
A	5.00 – 4.75
B	4.74 – 4.00
C	3.99 – 3.00
D	2.99 – 2.00

C. *Classification models*

Several supervised classifiers imbedded in the orange data mining software were used in this study. Each classifier was used to generate a models based on the features and the class value to answer the research question. The classifiers that were used are : KNN, Tree, SVM, Random Forest, Naïve Bays. Training and testing were conducted on the dataset available for this work. The dataset was divided into 66:34 for training and testing, respectively. To avoid overfitting and underfitting, 10-Fold cross validation was used. Fig.1 shows the workflow for academic performance of health science students' dataset.

D. *Model evaluaiton*

To evaluate the best fit model for this study, several metrics were used that include area under the curve (AUC), accuracy, F1, precision and recall. Students who are in need for consultation to enhance their academic performance can be assured by the best model evaluated by these measurements.

IV. RESULTS AND DISCUSSIONS

We evaluated 3468 student data records that covers a seven-year span between 2012-2019 to predict which students are likely to graduate with a "A" , "B", "C" or "D" cumulative GPA. We used five machine learning classifiers, KNN, Decision Tree, SVM, Random Forest, Naïve Bays to create the models. We show that the best performance based on AUC is random forest (.773) followed by naïve bayes (.758), SVM(.686), KNN(.684) and decision tree (.658) as shown in table 3. The confusion matrix for the random forest is another indicator of the usefulness of this algorithm (Fig. 2).

Further, to evaluate which features has more effect on predicting CGPA, three scoring methods were selected. Info. Gain, Gain ration and ReliefF. Preparatory year GPA ,Gender, and high school GPA were the top variables predicating student performance. Interestingly, standardized test did not predict graduating GPA, this is consistent with numerous studies world-wise that are revealing that standardized high school tests do not predict performance at college level [22].

Different research wok in machine learning in predicting students' performance has reported different results. Three studied have shown random forest classifier to outperform other classifiers. One study predicts students' academic performance based on demographics, student previous performance information, course and instructor information, and student general information, random forest was the most appropriate model [23]. A study to predict the right path for engineering students attending preparty year found random forest algorithm as the best fit model for the Mechanical engineering department [24]. Finally, random forest has outperformed other classifiers (decision trees, support vector machines, naive Bayes, bagged trees and boosted trees) in predicting students' final year performance based on information available at the end of the first year with 96% accuracy [25]. These results are consistent with our findings. However, two studies to predict students' GPA showed extreme gradient boosting classifier outperformed random forest [26]. And Naive Bayes scored 83.65%/ accuracy over random forest with 71.15% accuracy to show students' academic achievement at the end of the final year [27].

Information gained through this modeling can be used by universities for early intervention measures to support talented students who are good enough to gain admission yet struggle at the college level. This model is ready to be tested on batch of 2019-2022 once their information is available on the student's information system. Also, the data set can be expanded to include other tracks such Engineering, Sciences and Management, Arts and Education to be part of further studies to test the model.

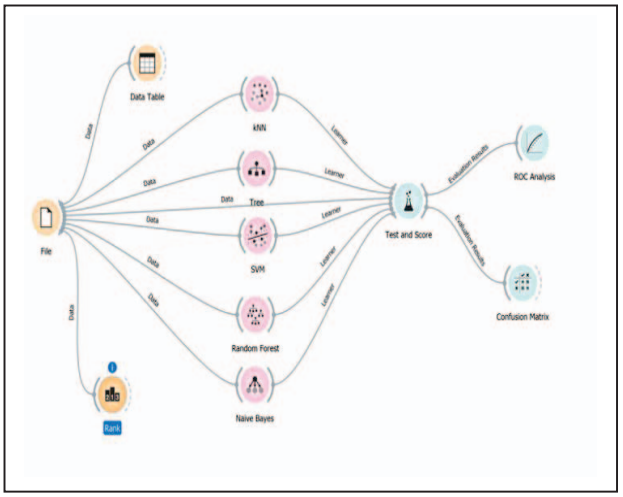


Figure 1. Orange workflow for health science students database

Table 3. Classifier's evaluation matrix

Model	AUC	CA	F1	Precision	Recall
kNN	0.684	0.662	0.639	0.630	0.662
Tree	0.658	0.659	0.654	0.652	0.659
SVM	0.686	0.645	0.604	0.611	0.645
Random Forest	0.773	0.705	0.693	0.690	0.705
Naive Bayes	0.758	0.660	0.661	0.663	0.660

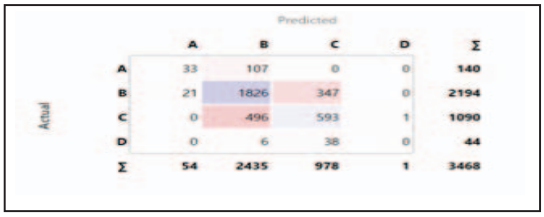


Figure 2. Random forest confusion matrix

V. CONCLUSION

The focus of this research was to apply multiple machine learning classifiers to predict students' performance in health colleges. Their performance is measured upon graduation with a "B", "C", or "D" GPA. Health college students should be motived to graduate with a B or higher GPA as they are excellent students who are good enough to get admission to health colleges but struggling at their university studies. Therefore, the outcome of this research is to serve a guidance to university staff to provide these students with the at most support they deserve.

ACKNOWLEDGMENT

I thank Mr. Yousef Salahat (IAU) for assistance with data collection and Ms. Wadhwa Aldossary for manuscript formatting .

REFERENCES

[1] K. Umamaheswari, S. Niraimathi,"A Study on Student Data Analysis Using Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3, Issue 8. August 2013.

[2] E. Osmanbegovic, M. Suljic, "Data Mining Approach for Predicting Student Performance", Journal of Economics and Business, Vol. X, Issue 1. May 2011.

[3] S. Rawat, S. Sreenatha et al., "Ascertaining the Factors Influencing Students' Performance for Engineering Pedagogy" Vol.28,pp. 30-33. January 2015.

[4] M. Suhaimi, N. Abdul-Rahman et al., " Review on Predicting Students' Graduation Time Using Machine Learning Algorithms", International journal of modern education and computer science, Vol. 11,issue 7. July 2019.

[5] G. Dekker, M. Pechenizkiy, M., " Predicting students drop out: A case study", Proceedings of the 2nd International Conference on Educational Data Mining, pp. 41-50, 2009.

[6] Tatar, S. Dilek, " Prediction of Academic Performance at Undergraduate Graduation: Course Grades or Grade Point Average?", Applied sciences., Vol. 10, Issue 14, July 2020.

[7] J. Superby, F. Meskens, " Determination of factors influencing the achievement of the first-year university students using data mining methods", the 8th international conference on intelligent tutoring systems, P.32: 234, 2006

[8] A. Shahiri, H. Wahidah, "A review on predicting student's performance using data mining techniques", Procedia mput Sci.,Vol.72, pp. 414–422, 2015.

[9] G. Zhang, M. Anderson et al., "Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study", Journal of Engineering Education, Vol. 93 (4), pp. 313–320, 2004.

[10] M. Goga, S. Kuyoro et al., "A recommender for improving the student academic performance", Social and Behavioural Sciences, Vol. 180 , pp. 1481 – 1488, 2015.

[11] M. Al-Barrak, M. Al-Razgan, "Predicting students final gpa using decision trees: a case study", International Journal of Information and Education Technology, vol. 6, no. 7, pp. 528, 2016 .

[12] Paris, L. Affecndy et al., "Improving Performance Prediction using Voting technique in data Mining", World Academy of Science, Engineering and Technology, Vol. 38, 2010

[13] A. Nandeshwar, S. Chaudhar, " Enrollment prediction models using data mining. Retrieved July 25, 2021, from http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf, (2009).

[14] D.Kabakchieva, "Predicting Student Performance by using Data Mining methods for classification." ,Cybernetics and Information Technologies, Vol. 13, 2013.

[15] R. Pereira, J. Zambrano, "Application of decision trees for detection of student dropout profiles In Machine Learning and Applications", IEEE International Conference, pp.528-531, IEEE. Decemebr , 2017.

[16] M. Quadri, N. V. Kalyankar. "Drop out feature of student data for academic performance using decision tree techniques." Global Journal of Computer Science and Technology, Vol.10, 2010.

[17] A. Purwinarko, W. Hardyanto, et al., " Academic achievement analysis of Universitas Negeri Semarang students using the naïve bayes classifier algorithm, Journal of Physics: Conference Series, Volume 1918, Mathematics and Its Application,2021.

[18] Z. Kovačić, "Early Prediction of Student Success:Mining Students Enrolment Data", Proceedings of Informing Science & IT Education Conference , 2010.

[19] Edin, and Mirza Suljić. "Data mining approach for predicting student performance." Economic Review 10.1 (2012): 3-12.

[20] N. Stanković1, M. Dragovan et al., "Artificial Neural Network Model for Prediction of Students' Success in Learning Programming", J. Sci.d, Res., Vol 80, March, 2021.

[21] D. Pranav, A. Ravina, et al., "Educational data mining for predicting students' academic performance using machine Learning", Materials today, proceedings, Vol 47, part 15, pp. 5260-5267, June 2021.

[22] E. Allensworth, k. Clark, "Are GPAs an Inconsistent Measure of College Readiness across High Schools? Examining Assumptions about Grades versus Standardized Test Scores", University of Chicago Consortium on School Research, April 2018.

[23] S. Amjad, M al-Emran, K,Shaalan, "Mining Student Information System Records to Predict Students' Academic Performance, March 2019,The International Conference on Advanced Machine Learning Technologies and Applications

[24] M. Ezz, A. Elshenawy, "Adaptive recommendation system using machine learning algorithms for predicting student's best academic program", 0202, Education and Information Technologies, 25:2733–2746

[25] V. Miguéis, A Freitas, P. Garcia, " Early segmentation of students according to their academic performance: A predictive modelling approach ", 2018, Decision Support Systems V. 115, Pages 36 - 51

[26] K. Jaiswal, P. Pathak, V. Pawar," Prediction of degree student achievement analysis",2021, IOP Conf. Ser.: Mater. Sci. Eng. 1070 012057

[27] Asif, R., Merceron, A., Abbas, S., & Ghani, N. (2017). Analyzing undergraduate students ' performance using educational data mining. Computers in Education, 113, 177–194.

Research Progress, Hot Spots and Prospects in the Field of Agricultural Innovation in China

——CiteSpace analysis based on 2258 articles

Shuqi Zhu，Qian Lu
School of Business, Guilin University of Electronic Technology
Guilin, People's Republic of China
zsq1024lucky@163.com

Yun Xiang
Lingnan College, Sun Yat-sen University
Guilin, People's Republic of China
xiangy63@mail.sysu.edu.cn

***Abstract*—Agricultural innovation is an important force driving the high-quality development of agriculture. Based on 2258 articles in China National Knowledge Infrastructure(CNKI) core journal database, and using Cite-Space software, we investigated the research progress, research hotspots and changes in the field of agricultural innovation in China from 2005 to 2020, trying to clarify the research trends and frontiers in the field of agricultural innovation. The results show that: ① The number of publications in the field of agricultural innovation in China presents an overall trend of "increasing volatility in the early stage, a short-term decline in the medium-term, and a rapid rebound in the near term", but the cooperative research between authors and institutions is relatively weak; ②The hotspots of agricultural innovation research focus on technology Innovation, institutional innovation, innovation evaluation and modern agriculture, etc., and show phase characteristics; ③The evolution of research hot topics in the field of agricultural innovation in China reflects a strong background of the times. Therefore, in the future, interdisciplinary research will be a new hotspot and direction in the field of agricultural innovation. It is urgent to carry out systematic research on theories, methods, and applications, and strengthen the research on the integration of agricultural innovation and practice.**

Keywords-agricultural innovation;Cite Space analysis; research progress;researchhotspots; research prospects

I. INTRODUCTION

Agricultural innovation is an important driving force for achieving high-quality agricultural development. At present, the main contradiction of China agricultural development has changed from insufficient total quantity to structural contradiction. However, it is a general consensus that the quality of China agricultural economy is low. It is urgent to improve the quality of agricultural economy. How to give full play to the driving role of agricultural innovation remains to be further explored.

It is generally believed that innovation factors have a positive impact on the high-quality development of the agricultural economy. However, traditional agricultural innovation research reviews mainly rely on reading a large number of documents and systematically reviewing them. The documents read are limited and inevitably subjective, and cannot clearly grasp the development process and relationship context of the field, and these complex relationships It can effectively catalyze the generation of new knowledge, while providing hotspot analysis and cutting-edge exploration directions, pointing out the research focus and direction of this field for scientific and technological workers, and improving scientific research efficiency. With the help of Cite-Space software for bibliometric analysis, these problems can be avoided. Based on this, this research attempts to use Cite-Space visualization software to provide a basis for follow-up research in the field of agricultural innovation in China.

II. DATA AND METHODS

A. Research tools and methods

This research uses Cite-space5.1.R6 version software to set the initial parameters, set the time span to 2005-2020, select the Node Type visualization node "Author", set the threshold to "TOP50", and time slice "per slice 1" ", the other options default, through the co-citation analysis theory and pathfinding network algorithm, respectively select the author (author), institution (institution) and keywords (keyword) as node types (node types) to generate a co-occurrence map, and then cooperate from the author systematic analysis of the development trend, key areas and research hotspots of agricultural innovation research in terms of network, institutional cooperation network, keyword network, keyword emergence, etc., in order to provide direction and reference for follow-up related research[1].

B. Data sources

Based on the CNKI core journal database of CNKI, search for subject terms such as "agricultural innovation", "agricultural technology innovation", and "agricultural innovation resources", the time range is 2005-2020, and a total of 2655 records were retrieved. However, due to the huge amount of literature related to the subject, the quality is uneven, and the degree of fit varies greatly. Therefore, the core journals were used as the source journals to refine the data, and the documents such as reports and book reviews were eliminated, and finally 2,258 valid documents were obtained, and then the document visualization analysis was carried out on this basis.

III. ANALYSIS OF RESEARCH PROGRESS IN THE FIELD OF AGRICULTURAL INNOVATION

A. Analysis of the number of articles issued

The annual statistical analysis of the number of articles and the distribution of topics can intuitively reflect the degree of importance the academic community attaches to a certain research field, reveal the development status, and

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00016

predict the research prospects of the agricultural innovation field [2]. Generally speaking, the more articles are published, the more active the research in this field is. Through the time distribution of relevant literature in the field of agricultural innovation research, we can basically understand the degree of domestic research activity in the field of agricultural innovation (see Figure 1). In general, since 2005, the number of publications in the field of agricultural innovation in China has shown a trend of "increasing volatility in the early period, a short-term decline in the mid-term, and a rapid rebound in the near term", basically showing the "N-type" characteristics. Among them, the number of articles issued increased from 77 in 2005 to 166 in 2020, an increase of 115.58%. The average amount of articles issued from 2005 to 2020 was 150, indicating that the field of agricultural innovation has maintained a high degree of research enthusiasm, and after 2018 ,the number of publications has shown a rapid upward trend, rising from 98 in 2018 to 166 in 2020, an increase of 69.39%, which further reflects that the current agricultural innovation field is still receiving strong attention and will remain one of the important research focuses in the future.

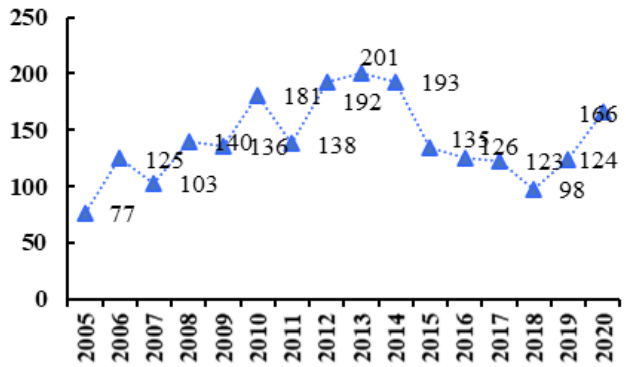


Figure 1. Number of research papers issued in the field of agricultural innovation in China from 2005 to 2020.

B. *Analysis of Author's Collaboration Network*

Prolific authors in the author cooperation network refer to the most active and fruitful scholars in this field, as well as those who have made important contributions to the research in this field within a certain period of time [3]. It is generally believed that the author's name in the knowledge graph and the size of the node circle are proportional to the amount of author's research on the topic. The number of nodes represents the number of core authors under the selected parameter combination, and the number of connections represents the number of authors between authors. The degree of cooperation, the number of connections and the closeness of the cooperative relationship show a positive relationship [4]. A map of authors' cooperation in the field of agricultural innovation research in my country since 2005 was drawn (see Figure 2), and a total of 225 nodes and 178 connections were found, with a network density of 0.0071. The number of connections is less than the number of nodes, indicating that authors in the field of agricultural innovation The cooperative relationship between them is relatively weak, and only a small number of relatively stable cooperative teams have been formed, and there is a lack of effective cooperation mechanisms and leading figures. The cooperative research between the authors needs to be improved.

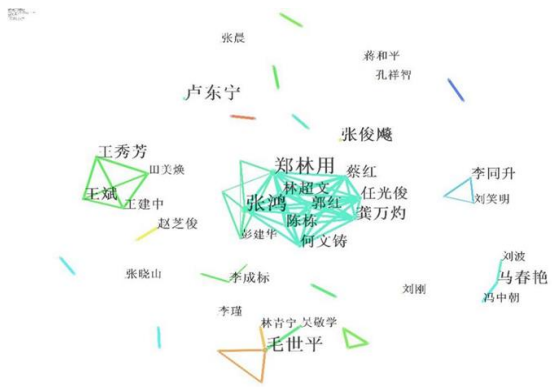


Figure 2. Authors' cooperation map in the field of agricultural innovation in China from 2005 to 2020.

Since each node in the graph represents an author, by observing the size of the node and the thickness of the connection, the amount of articles published by the authors and the intensity of cooperation between authors can be reflected [5]. Further analysis of Figure 2 shows that the cooperation network between authors has two major characteristics: First, judging by the number of posts and the size of the nodes, since 2005, there have been a number of agricultural innovation research fields such as Zhang Hong, Zheng Linyong, Mao Shiping, and Zhang Junbiao. The core authors represented by Ma Chunyan, Ma Chunyan and Ren Guangjun have found that there are relatively few author teams with a certain degree of cooperation. Second, the cooperative relationship among authors in the field of agricultural innovation generally shows the characteristics of "overall dispersion and partial concentration". Most authors are in a state of independent research, and only a few authors have established cooperative relations, and mainly single-line cooperation. It reflects the urgent need to strengthen cooperative research among scholars in the field of agricultural innovation.

C. *Analysis of the cooperation network between research institutions*

The analysis of core academic institutions helps to grasp the distribution of academic resources at a macro level, and is conducive to the targeted supply and optimal allocation of academic resources. Use Cite-space software to obtain a co-occurrence map of research institutions (see Figure 3), where the number of nodes represents the number of core authors proposed under a specific parameter combination, and the number of connections represents the degree of cooperation between institutions under the subject. In the knowledge map the size of the institution's name and the size of the node circle are directly proportional to the amount of research published by the author on the topic. The map has 161 nodes, 38 connections, and a network density of 0.003, indicating that there are many research institutions in the field of agricultural innovation in my country, but the cooperation between research institutions is still relatively weak.

Figure 3. China's cooperation map of agricultural innovation research institutions from 2005 to 2020.

Note: Due to space reasons, only the organizations mentioned in the text are listed:

中国农业科学院农业经济与发展研究所: the Institute of Agricultural Economics and Development of the Chinese Academy of Agricultural Sciences;华中农业大学经济管理学院: the School of Economics and Management of Huazhong Agricultural University;中国人民大学农业与农村发展学院: the School of Agriculture and Rural Development of Renmin University of China

Further analysis of figure 3 shows that colleges and universities and scientific research institutions are the main research forces. Among them, the Institute of Agricultural Economics and Development of the Chinese Academy of Agricultural Sciences, the School of Economics and Management of Huazhong Agricultural University and the School of Agriculture and Rural Development of Renmin University of China are the most prominent. Research institutions represented by three institutions have made great contributions to agricultural research. The research directions of these research institutions are all agricultural economics, and they have played a leading role in agricultural economic research in the field of agricultural innovation in my country. Relatively speaking, other the distribution of research institutions is relatively scattered, indicating that the exchanges and cooperation between institutions are relatively weak. In the future, the field of agricultural innovation should increase cooperation between institutions. Collaborative papers can better bring together the wisdom of different scholars or the research characteristics of different research units. The cooperation will form a more complex knowledge structure, a more diverse range of topics, and more. Cross-cutting research perspectives, more multidisciplinary theoretical models and research methods, etc., are all conducive to improving the quality of collaborative papers [6].

IV. RESEARCH HOTSPOTS IN THE FIELD OF AGRICULTURAL INNOVATION AND ANALYSIS OF THEIR EVOLUTION

A. High-frequency keywords

In a certain field, if one or several keywords are found repeatedly in numerous documents, then the topics represented by these keywords are the research hotspots in this field [7]. Use Cite-space software to obtain 2258 key word co-occurrence maps in the field of agricultural innovation (see Figure 4), obtain a total of 364 nodes, 731 connections, network density 0.0611, index Q = 0.8798, greater than 0. 3. S = 0.5555, greater than 0.5, indicating that the clustering result is reasonable. Further analysis believes that the number of connections is far greater than the number of nodes, reflecting the relatively close relationship between the keywords of the research topics, and the formation of natural clustering of keywords represented by technological innovation, modern agriculture, agricultural innovation, technological innovation, etc. To a certain extent, reflects the research focus and hot spots in the field of agricultural innovation in my country. From the point of view of the network density of 0.0611, although many researches have been carried out on agricultural innovation, the depth of research still needs to be further explored.

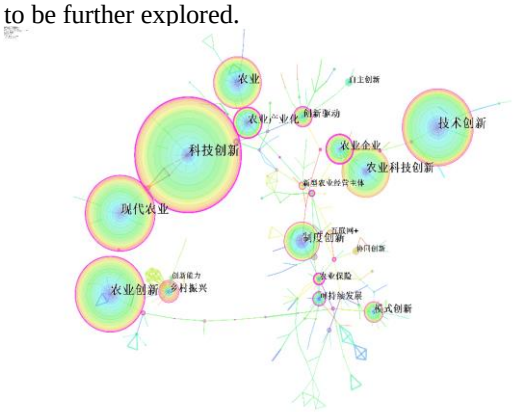


Figure 4. A map of research keywords in the field of agricultural innovation in China from 2005 to 2020.

Note：Chinese and English are as follows

科技创新: scientific and technological innovation; 现代农业: modern agriculture, 农业创新: agricultural innovation, 农业产业化: agricultural industrialization; 农业企业: agricultural enterprises; 创新驱动: innovation drivers; 农业保险 agricultural insurance; 技术创新: technological innovation; 乡村振兴 Rural revitalization；可持续发展 sustainable development；制度创新 institutional innovation；互联网+ Internet +；可持续发展 sustainable development；协同创新 collaborative innovation；新型农业经营主体 new agricultural business entities；自主创新 independent innovation.

Analyzing the size and color depth of each node in the keyword map of the agricultural innovation research field, it is concluded that the nodes of agricultural scientific and technological innovation are much larger than other nodes, indicating that scientific and technological innovation plays an important role in the field of agricultural innovation[8]. In addition, the outer circles of scientific and technological innovation, technological innovation, modern agriculture, agricultural innovation, agricultural industrialization, agricultural enterprises and innovation drivers, and agricultural insurance are all in red, indicating that these keywords have received continuous attention in recent years. On this basis, the cluster analysis of keywords can get five larger clusters of "agricultural technology innovation", "agricultural system innovation", "agricultural modernization", "agricultural industrialization" and "agricultural bio-industry". The category label reflects that these five themes are important research hotspots in the field of agricultural innovation research.

B. Changes in research topics

According to the Cite-space software, the keyword co-occurrence time zone network mutation analysis diagram can be drawn, and you can find the changes in research topics in a certain field and judge the research trend in a certain field [9]. The general experience believes that the darker the color of the node in the network mutation

analysis graph of the keyword co-occurrence time zone, the more active the corresponding keyword is in the near future [10]. Therefore, by drawing the Time-zone map of the key words in the field of agricultural innovation research (see Figure 5), the evolution trend of hot topics in agricultural innovation research can be determined.

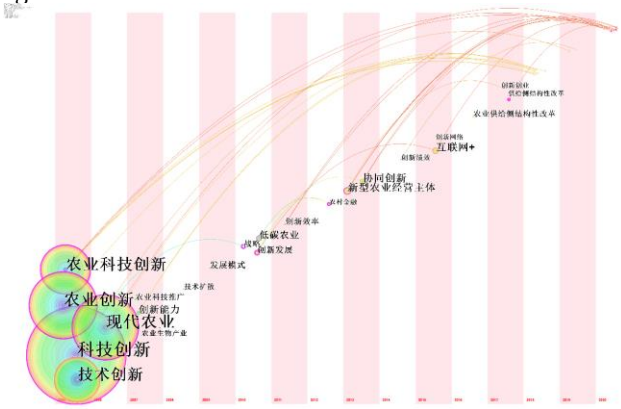


Figure 5. Time-zone map of keywords in the field of agricultural innovation research in China from 2005 to 2020.

Note: Chinese and English are as follows

农业科技创新:agricultural scientific and technological innovation;农业创新:agricultural innovation;现代农业:modern agriculture;科技创新:scientific and technological innovation;技术创新:technological innovation;低碳农业:low-carbon agriculture;农业金融:agricultural finance;新型农业经营主体:new agricultural business entities;协同创新:collaborative innovation;互联网+ :Internet +;农业供给侧结构性改革:agricultural supply-side structural reform;技术扩散:technology diffusion;农业生物产量:agricultural biological output;发展模式:development model;创新效率:Innovation efficiency;创新绩效:Innovation performance;创新网络:Innovation network.

On the basis of analyzing figure 5 and studying related literature, it can be found that the changes in research hotspots in the field of agricultural innovation mainly show the following characteristics: First, in 2005, the amount of research literature related to popular topics was large and high-quality, such as "Science and Technology Innovation" , "Technical Innovation", "Modern Agriculture", etc., have laid a solid foundation for the field of agricultural innovation, and these keywords have continued to this day, indicating that the hot topics in the field of agricultural innovation have a strong continuity. Second, the hot research in the field of agricultural innovation reflects strong characteristics of the times, and is closely related to the country's agricultural policies. "Agriculture", "ecological circular agriculture", etc., after 2015, "Internet+", "innovation network", and "rural supply-side structural reform" have become high-frequency keywords. Third, in recent years, although keywords worthy of attention have appeared, due to the short research time, the performance of the research results in the time-zone is not obvious, and further analysis is needed through keyword highlighting.

V. RESEARCH CONCLUSIONS AND PROSPECTS

A. *Conclusions*

First, the number of publications in the field of agricultural innovation in China shows an overall trend of "increasing volatility in the early period, a short-term decline in the mid-term, and a rapid rebound in the near term", but the cooperative research between authors and institutions is relatively weak..

Second, the research hotspots of agricultural innovation focus on technological innovation, institutional innovation, innovation evaluation, and modern agriculture, etc., and they show phased characteristics. The five themes of agricultural technology innovation, agricultural system innovation, agricultural modernization, agricultural industrialization and agricultural bio-industry are important research hotspots in the field of agricultural innovation research.

B. *Research outlook*

At present, China's agricultural innovation work is still advancing continuously.

The systematic research on theories, methods and applications in the field of agricultural innovation urgently needs to be further improved. Research in the field of agricultural innovation focuses more on "what is agricultural innovation" and "how to achieve agricultural innovation". There is relatively little research on "why agricultural innovation". For example, "how agricultural innovation can achieve rural revitalization under the new development pattern" "How does the quality of agricultural innovation promote the high-quality development of agricultural economy", in-depth exploration of these issues is not enough, and systematic exploration from multiple dimensions such as theoretical origin, method verification, and application practice is needed.

REFERENCES

[1] Chen Chaomei, Chen Yue, Hou Jianhua, et al. Cite Space Ⅱ: Identification and visualization of new trends and new developments in scientific literature[J]. Journal of Information, 2009.

[2] Yang Ming, Ye Qing, Jin Jing. Bibliometric analysis of agricultural development research from 1900 to 2010 [J]. Theory and Modernization, 2012, 000(004): 120-126.

[3] Geng Liuli, Li Na, Wang Qi, Jin Wei. Analysis of my country's Rural Industrial Convergence Science Knowledge Graph——Based on CNKI Core Journals and CSSCI Data from 2014 to 2018 [J]. Journal of Xichang College (Social Science Edition), 2019, 31(04):67-72+113.

[4] Liu Shuai, Junhong Wu. Visualization Analysis of Domestic College Counselors in Recent 20 Years Based on Cite Space[J]. International Journal of Higher Education Teaching Theory,2020,1(4).

[5] Zhang Hong, et al. The construction framework and operating mechanism of the Sichuan innovation team in the national modern agricultural industrial technology system[J]. Science and Technology Management Research, 2010, 30(09): 70-72+84.

[6] Mao Shiping, Yang Yanli, Lin Qingning. The evolution and effect evaluation of my country's agricultural science and technology innovation policies since the reform and opening up: empirical evidence from my country's agricultural scientific research institutions[J]. Issues in Agricultural Economics, 2019(01): 73-85.

[7] Wang Bin, et al. On the construction of the theoretical system of agricultural science and technology venture capital [J]. Jiangsu Agricultural Sciences, 2014, 42(04):387-391.

[8] Wang Xiufang, Yu Shusheng, Sun Yujie. Research on the External Environmental Issues of Agricultural Structural Adjustment[J]. Issues of Agricultural Economics, 2002(11): 16-19.

[9] Tang Yun Xu,Hai Ni Qu,San Shan Zhao,Mei Xia Zhang,Xiu Yang,Xu Yan Huang. The Visualization Analysis of Research Hotspot and Frontier Technology of the Smart Power Distribution and Utilization Based on the Cite Space[J]. Energy and Power Engineering,2017,9(4B).

Chen Peng, Tian Yang. Analysis of Research Hot Spot and Trends of Bioinformatics Based on Cite Space[J]. Nanoscience and Nanotechnology Letters,2018,10(11).

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Gender Classification for Online Shopping

Yimeng Zhang
Taiyuan No. 20 High School
Taiyuan, China
E-mail: snowblanche@sina.com

Abstract— **The proliferation of Online shopping has been increasing in the past decades. Different online shopping companies investigate on precise shopping recommendation system based on the customers online viewing log and purchase log data. Even though the online shopping recommendation has been investigated for several years, both industrial and academia could not propose a generalized and efficient model to predict customers' shopping demand. Recently, the customers' gender information attract people's attention since the gender information reflects the customers' shopping behavior and preference. Nevertheless, the gender information collected from online shopping system are neither intact nor fake since customers don't want to leak their privacy. Hence, the estimation of customers' gender becomes critical for the online shopping recommendation system. This paper focuses on gender estimation based on customers' online viewing log collected by the FTP group, a leading information and communication enterprise in Vietnam. Given the imbalanced (population of female is 3 times of male) and ambiguous data, we propose our approach to estimate the gender with 75% accuracy. Specifically, we observe that the female samples naturally form 3 clusters when we select duration of session, number of items viewed, and average time spent on each item as the features. Then, we naturally divide the female set into 3 subsets and merge them with male set to generate the 3 training sets, which don't have imbalance issue. 3 individual models are trained from these 3 training sets and a new classifier is used to make the final decision based on the output of these 3 models. Our experimental results show that we can achieve 75% accuracy while the running time is less than 7 seconds.**

Data Mining; PAKDD Contest; Classifier; Cluster; Imbalance.

I. INTRODUCTION

Online shopping has already become a part of our daily life. Tons of online shopping information including discount, sale and advertisement are sent to our email as well post mailbox every day. The competition among different online shopping companies is getting more and more severe. Hence, the companies providing more convenient shopping experience to customers would become the winner. Hence, these companies spend a lot of money on investigating the customers' shopping behavior and preference to make a better shopping recommendation system, which can provide customers the most needed items. Nevertheless, these companies realize the customers' gender information is very critical in the online shopping recommendation system [1].

Then, a natural question is popped up: can we get customers' gender information? Unfortunately, the answer is no [2]. The reason is easy to understand. We all have the experience to register on some website, where we are asked to select our gender. However, most people ignore that option or randomly select the gender. People don't care about the gender information because (i) they don't want to leak their privacy to the website since they might just purchase the items on this website once [3]. (ii), they don't want to waste time on make the selection. hence, the gender information collected from the website is far less than enough to contribute to the online shopping recommendation system [4]. Therefore, the estimation of customers' gender becomes necessary.

This paper proposes an approach to estimate the gender of customers shopping on FTP group, a leading information and communication enterprise in Vietnam. FPT runs several B2B2C (business-to-business-to-customer) services that provide online shopping sites and mobile applications for small and medium sellers. Transaction data, such as product browsing and purchasing activities, from buyer, and product portfolio, from seller, can be aggregated, to provide more efficient buying and selling experiences. Data mining techniques are applied to predict the optimal organization and display of products that maximize the chance of bringing useful information to user, facilitate the online purchases. Given the online viewing log, we estimate the customers' gender for the FTP group to make a better online recommendation system.

The data we have from FTP group is demonstrated in Table 1, which shows 2 samples in the training data set. Session ID is the ID of each view session while "start" and "end" represent the start time and end time of this session. The "Viewed Items" stores all the items the user views in this session and the "Class Label" is the corresponding gender. More specifically, items are classified to 4 categories: A represent the most generalized items and those starting with 'D' correspond to individual products. The IDs which start with 'B' and 'C' are associated with subcategories and sub-subcategories, respectively. The format of test data set is as same as training data set except for the class label. Both test and training data set contain 15000 samples. we get these data from PAKDD 2015 contest website. The size of training and testing file are 224 and 223KB respectively. We also get the correct class label for the test set. Then, it is very convenient for us to calculate the accuracy of our own result and compare our approach with other public results in the PAKDD 2015 contest ranking list.

As we can see from the Table 1, the relationship between 4 columns and class label are too loose, we need to extract meaningful feature from the data set. The difficulty here is how to define the appropriate features that provides the maximum estimation accuracy. This requires the deep understanding of online shopping

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00017

Session Id	Start	End	Viewed Items	Class Label
u10001	11/14/2014 0:02	11/14/2014 0:02	A00001/B00001/C00001/D00001/	Female
u10002	12/12/2014 14:12	12/12/2014 14:12	A00002/B00002/C00002/D24897/	Male

Table 1, Data format.

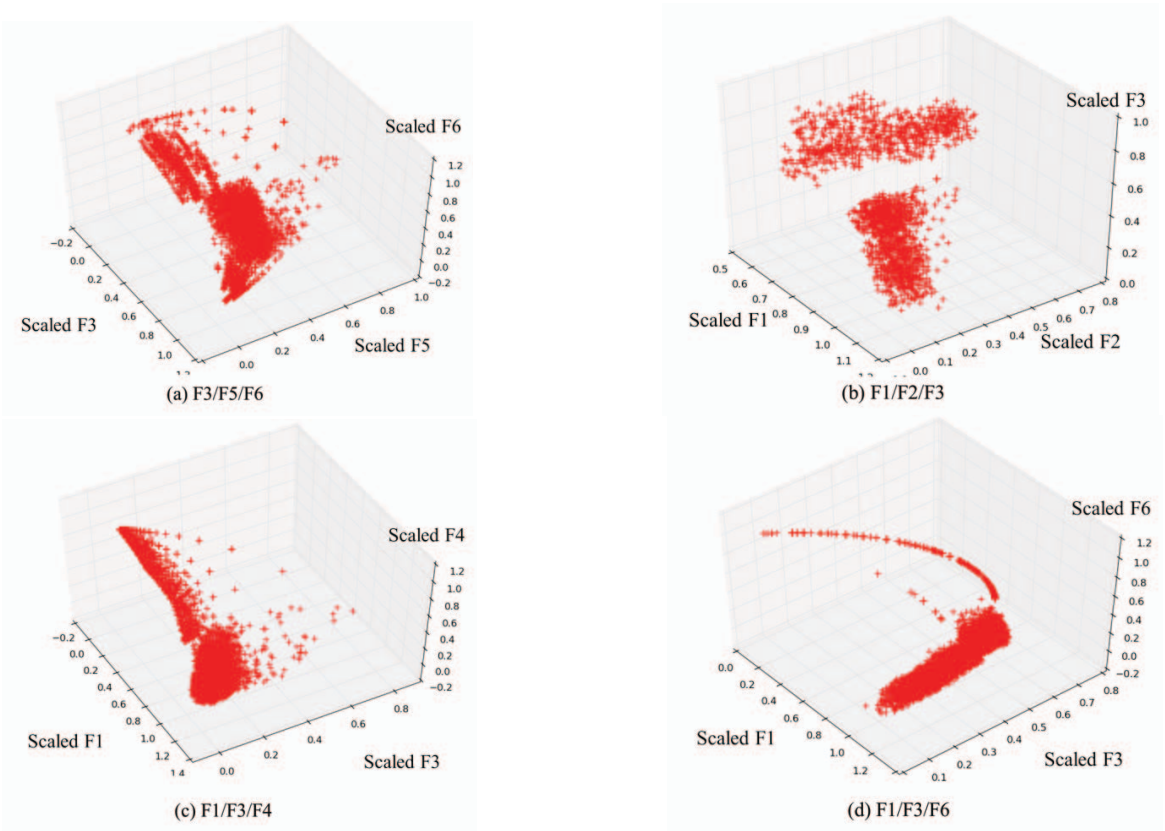


Figure 1, Data visualization for different feature combinations.

behavior and psychologic analysis. Another issue in the dataset is imbalance issue. Training set includes 11703 Female and 3297 Male samples. This imbalanced training dataset causes poor estimation result, which would be shown in section 2.1.

Even though the population of female samples is much larger than the population of male samples, we know that females have different personalities. For example, I am a girl while my personality is not like a typical girl. On contrast, my personality is close to a boy. Then, we can divide the female set into several (3) subsets according to personality. By selecting the feature, the subsets have the similar population which is almost equal to number of male samples. Hence, the imbalance issue is solved. The details are demonstrated in section 2.2 and 2.3.

Our contribution has two folds:

- We extract 6 features from the online viewing log. By applying data visualization on different feature combinations, we select 10 combinations as the feature candidates. We take advantage of personality diversity to further filter out the feature combination candidates and divide the female training set into 3 subsets. Then, we generate 3 balanced sub training sets. Basically, we address the imbalance issue in this dataset.
- We the proposed approach with different classifiers and achieve 75% accuracy while the running time is as short as 7 seconds.

II. DESIGN

In this section, we demonstrate our approach in detail.

A. Feature Definition

We use the sk to denote the k-th session. The start and end of k-th session are represented by ts(sk) and te(sk). The corresponding items are stored in the item set: Item(sk)={ item(sk)1, item(sk)2,… item(sk)i,…}. We have 6 candidate features and their definitions are as follows:

Definition 1: duration of session. Duration of the k-th session, Du(sk)= te(sk)- ts(sk).

Definition 2: number of items viewed. The number of items viewed in the k-th session, N(sk)= | Item(sk)|.

Definition 3: average time spent on each item. The average time spent on each item in the k-th session, Tavg(sk)= Du(sk)/ N(sk).

Definition 4: start time of the session. The time the session is opened, T_s(sk)= ts(sk)

Definition 5: Number of the most generalized (‘A’ category) items. NA(sk)= |{ item(sk)j| item(sk)j starts with “A”}|

Definition 6: Number of the most individual ('D' category) items. ND(sk)= |{ item(sk)j| item(sk)j starts with "D"}|

Let us use an example to show the intuition of these features. As a male, I usually have a list before I shop online. Then, when I view the items, I prefer staying on each webpage for a while because I would like to get more information. So, I think the average time spent on each item for male might be longer than female. The start time of the session maybe important too. Because I would never touch the shopping websites during the daytime. The number of most generalized/individual items shows the customers' shopping behavior and preference.

	F3/F5/F6	F1/F2/F3	F1/F3/F4	F1/F3/F6	F1/F3/F5
Accuracy	0.65	0.63	0.61	0.55	0.54
	F2/F3	F2/F3/F4	F2/F4/F6	F1/F2/F4	F1/F4/F5
Accuracy	0.54	0.53	0.51	0.51	0.50

Table 2, accuracy for different feature combination.

B. *Feature Selection*

Given the 6 featured defined in section 2.1, we need to decide which features are used in our approach. At first, we run the random forest classifier on the combination of these 6 features, which means we need to test 26=64 combinations. Since the space is limited, we show the accuracy of top 10 combinations in table 2. Since these 10 feature combinations provide more than 50% accuracy, all of them are candidates.

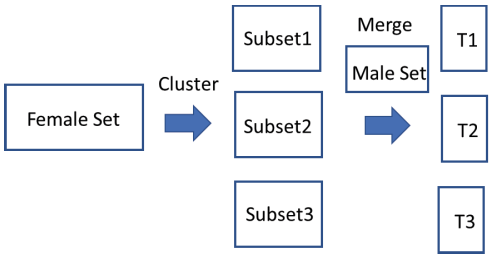


Figure 2, Training set generation.

C. *Clustering based Data Set Cut*

We know that the diversity of personalities exists in all the communities. The female samples should also reflect the personality diversity. We assume the diversity naturally forms several clusters in the female samples. In order to verify this assumption, we plot the female samples defined by top 4 feature combinations in figure 1.

Fortunately, the second highest accuracy combination, i.e., F1/F2/F3, depicts the clear 3 clusters. Since we know the accuracy of these combination is affected by the imbalanced samples, the highest accuracy combination without showing the personality diversity should not be the appropriate the feature combination. Therefore, we select the feature combination F1/F2/F3 as our feature, which shows the clear personality diversity.

Then, we use k-means method to find the 3 clusters. The number of samples in these 3 clusters are 3108, 4925 and 3670 respectively. They are almost balance. Then, as shown in figure 2, we generate 3 training set by merging these 3 female clusters (subsets) with male set. We cut the female data set by clustering method, which smartly solves the imbalance issue.

D. *Training Model*

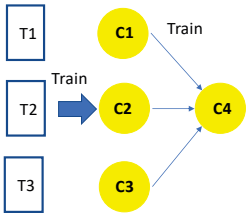


Figure 3, Training Model.

Given the training set generated in section 2.3, we demonstrate our model in this section. As figure 3 shows, we train the 3 model individually on each training set. Then we get classifier C1, C2 and C3. When we estimate the gender for a sample in test set, we input this sample into 3 classifier and get 3 corresponding results. For example, the results are "Female", "Male" and "Male" for C1, C2 and C3 respectively. Then, the question is which one is correct and how to make the final decision? The basic idea is to vote on these 3 results. However, using majority voting doesn't provide any sense. Alternatively, we use another classifier, C4, to smartly make the final decision. We use the output of C1, C2 and C3 as the input and use the true class label as the output to train C4.

The intuition of this approach is the personality of all the females are different and can be classified into several classes. Then, we use these subclasses to generate different training model. When we estimate the gender for the sample in test set, we firstly decide the gender for different female personality. For example, if the sample is a female sample, this stage would decide which personality it belongs to. If this sample is a male sample, this stage would provide the correct result with high confidence. Hence, we decide the personality of this sample first and we decide the gender at the second stage (using C4).

E. *Putting Together*

The whole approach consists of the 3 following steps:

• Step 1, Data cleaning and Feature extraction. Since we select F1/F2/F3 combination as our feature, we read the training set file and check if there exist any samples doesn't have intact information. In other words, we check if some sample doesn't have all the attributes. We find that there are 172 samples don't have all the 4 attributes. This is very normal in the practical data set. Since the number of the incomplete samples (172) is much smaller than the whole number of samples in training set, we can just remove these samples from the training set. Then, F1, F2 and F3 are calculated as their definitions in section 2.1.

• Step 2. We use K-means to find the 3 clusters in female set according to the personality diversity. Then, we merge these 3 clusters with the male set to generate the 3 training sets t1, t2 and T3.

• Step 3. We train 3 models, C1, C2 and C3 on 3 training sets individually. Since our approach is independent of the classifiers, we may use SVM, random forest, decision tree as the classifier. We show the accuracy of the different combination in the section 3. For example, we set C1 to be SVM, C2 to be random forest, C3 to be SVM and C4 to be decision tree. Then, we train the 4 classifiers and compare the result.

Eventually, we select the combination of classifiers with the highest accuracy as our final approach.

III. EVALUATION

We demonstrate the evaluation of our training model in this section. The ground true is already published at PAKDD 2015 competition website. Thus, we know the class label of test data set. Since the data set is imbalanced, we use the "balanced accuracy" to evaluate our approach. The definition of balanced accuracy is BAC (predict, gender)=(ACCm(predict,gender)+ACCf (predict,gender))/2, where ACCm ¬and ACCf are the accuracy of predicted male and female respectively. "predict" and "gender" represent our predicted result and ground truth. The formal definition of ACCm and ACCf are as follows:

$$ACC_m(predict, gender) = \frac{|j: predict_j = genrder_j = Male|}{|j: genrder_j = Male|}$$

$$ACC_f(predict, gender) = \frac{|j: predict_j = genrder_j = Female|}{|j: genrder_j = Female|}$$

A. Feature Selection

We select clustering features among the six features that most closely match the sample numbers of the female and male sets. F1 / F2 / F3 can classify a female sample set into three groups with the most similar number of samples. Through F1, F2, and F3, clusters 1,2,3 have 3108, 4925, and 3670 samples, respectively. On the other hand, the other feature combinations are more biased in a particular cluster, resulting in a large number of samples

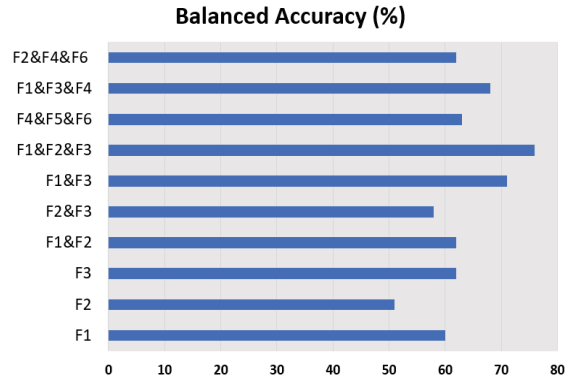


Figure 4, Balanced Accuracy with different feature set. The accuracy depends on the feature set.

B. Balanced Accuracy

We achieve 75% accuracy using Feature 1, 2, and 3. We set the training and test set from PAKDD as ground truth and verify the prediction results of our model. When verifying the accuracy of our model, we measure the accuracy based on the balanced accuracy. This is because our dataset is imbalanced, so if you do it by measuring the existing accuracy, you will get a bias result on the female set. To avoid this, we measure the accuracy through balanced accuracy. Basically, the balanced accuracy measures the predicted results for male and female, and then averages them.

C. Computational Overhead (Running Time)

In order to apply our approach into practical and commercial scenario, the computational overhead is a critical factor since our approach needed to be run for thousands of times on the large-scale data sets. In addition, the running time in this experiment include both feature extraction and training process. We know the online shopping data is dynamically changing every day or every second. If the company uses our approach for the different data collected at different time, the time spent on feature extraction should be considered. Hence, we evaluate the overhead (running time) of different classifier and feature combination. The corresponding results are summarized in table 3.

	Random Forest	SVM	Decision Tree	Gaussia nNB
F1/F2/F3	5s	6s	8s	6s
F4/F5/F6	6s	7s	7s	5s
F1/F3/F4	5s	8s	9s	4s
F1/F3/F6	5s	7s	12s	5s
F1/F3/F5	5s	7s	8s	6s
F2/F3/F4	6s	6s	7s	6s
F1/F2/F4	5s	9s	8s	7s
F1/F4/F5	6s	8s	9s	6s

Table 3, overhead for different classifiers.

The trend in this result is: the larger number of features we use, the more time is consumed. This is reasonable since the feature extraction involves I/Q operation and the more features we need extract, the more time we need to process the file. Random forest consumes the least time because we can control the depth of the tree in the random forest algorithm while the accuracy is not sacrificed. Even though the running time is important, the accuracy is more important. If the approach could not give us a good result, less running time is meaningless.

IV. CONCLUSION

This paper focuses on mining the gender information from the online shopping viewing log provided by FTP company. Given the meaningless raw attributes, we extract 6 meaningful features from the data set. In order to overcome the Female/Male imbalance issue, we take advantage of personality diversity to divide the female set into 3 subsets with similar size. Then, a 2-layer classifiers network is applied to estimate the gender. Our evaluation results show that our model of gender prediction achieves balanced accuracy of 75% while running time is 7 seconds. Even though the winner in PAKDD 2015 contest achieved 87% balanced accuracy, our approach is more light weight and efficient because the winner's method is too complicated.

REFERENCES

[1] Tomson Reuters, Article Title, https://blogs.thomsonreuters.com/answerson/ business-case-gender-parity/.

[2] Van Slyke, Craig, Christie L. Comunale, and France Belanger. "Gender differences in perceptions of web-based shopping." Communications of the ACM 45.8 (2002): 82-86.

[3] Lin, Xiaolin, et al. "Exploring Gender Differences in Online Consumer Purchase Decision Making: An Online Product Presentation Perspective." Information Systems Frontiers (2018): 1-15.

[4] Kim KJ, Ahn H. A recommender system using GA K-means clustering in an online shopping market. Expert systems with applications. 2008 Feb 1;34(2):1200-9

Mining model of purchase intention based on support vector machine algorithm

Mengtong Zhang
School of International Education
Wuhan University of Technology
Wuhan,China
zhang_000.t@whut.edu.cn

Jichang Dong
School of Science
Wuhan University of Technology
Wuhan,China
2504157235@qq.com

Wenkan Huang
School of Computer Science and Artificial Intelligence
Wuhan University of Technology
Wuhan,China
769397149@qq.com

***Abstract*—The automobile industry is an important pillar industry of the national economy, while new energy automobile industry is a strategic emerging industry. Firstly, the Spearman correlation coefficient of each index of each brand is calculated, and then the influencing factors of each brand are obtained, and the customer mining model of each brand is established. Finally, it tests the willingness of 15 target customers to buy electric vehicles and conducts k-fold cross-verification. The results show that the purchase user's number of different brands is NO.2 and 4 for brand 1, NO.7 for brand 2 and NO.12 for brand 3. The comprehensive prediction accuracy is as high as 0.94, indicating that the model has high sensitivity and the test results are accurate.**

Keywords-New energy automobile; Spearman correlation coefficient; SVM support vector machine model; K-fold cross-examination; Customer mining model

I. INTRODUCTION

The automobile industry is an important pillar industry of the national economy, and the new energy vehicle industry is a strategic emerging industry. Vigorously developing new energy vehicles represented by electric vehicles is an effective way to solve energy and environmental problems, with broad market prospects. However, for electric vehicles, as an emerging thing, consumers still have some doubts in some areas compared with traditional vehicles, and their market sales require scientific decision-making[1-2].There are many influencing factors that determine whether the target customer is going to buy electric vehicles, including the factors of the electric vehicles themselves and the personal characteristics of the target customer. Therefore, quantitative statistics on the data are conducted[3-4](The data in this paper comes from: http://shumo.neepu.edu.cn/).

II. ANALYSIS OF THE FACTORS INFLUENCING THE SALES OF THREE BRANDS OF ELECTRIC VEHICLES

By observing the data, it can be learned that its problems basically revolve around the "economy", "region" and "number of families" of car buyers. According to be seen from 0 / 1 data, this is a second classification problem. This article can divide the car buyers according to different brands and analyze the correlation of a1－a8, B1－ B17 and purchase willingness respectively. The stronger the correlation, the greater the impact, so as to determine the impact of a given factor on the purchase behavior. The solution idea is shown in Figure 1.

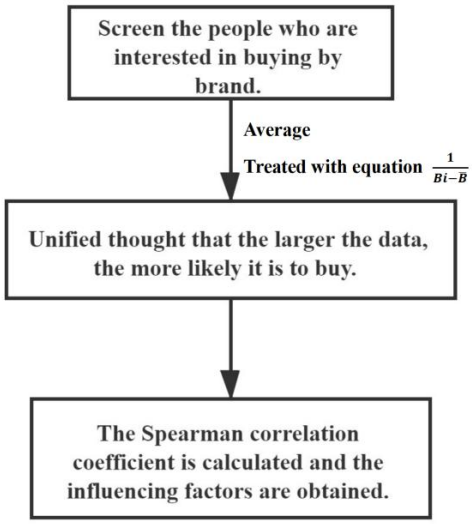


Figure 1. Thought diagram

For the cleaned data, all users willing to buy electric vehicles are selected, and then those willing to buy will be divided according to different brands. Learn through the information:

- If the data quantity is large enough, according to the large number law, the central limit theorem can guarantee an approximate normal distribution;
- If the continuous data meet the normal distribution and meet the linear relationship, the pearson correlation coefficient is the most appropriate.

If any of the above conditions are not met, the spearman correlation coefficient is used. In this paper, the data were analyzed with SPSS(Statistical Product and Service Solutions) to obtain linear regression analysis of 25 factors of three brands, among which the factor significance level is below 0.05 after regression, so there is a reason to think that it does not conform to the normal distribution and its data is discrete, so that the spearman correlation coefficient was selected for testing.

The analysis shows that the greater the B2, B3, B5, B7, B9, B10, B13, B14, B15 data, the more likely the people to buy electric vehicles; the closer the B1, B4, B6, B8, B11 data to the median, the more likely to buy the electric

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00018

vehicle; the smaller the B12, B16, B17 problem corresponding data, the more likely it is to buy an electric vehicle. For this, process the data:

- For B1, with only 1,2,3, changing 3 to 1 makes the greater the value, the greater the willingness to buy.
- For other factors, the paper calculates the average of one problem of the people willing to buy, thinking that the closer to this average, the more likely to buy.

Use the following formula for analyze:

$$\frac{1}{|\mathrm{Bi}-\overline{\mathrm{B}}|}(\mathrm{I}=1,4,6,8,11) \quad (1)$$

The spearman correlation coefficient is calculated, and the influencing factors of the three brands are shown in Table 1.

TABLE I. TABLE TYPE STYLES

Brand name	Influencing factor
Brand name 1	a1-a8，B16，B17
Brand name 2	a1-a8，B12，B13，B15，B16，B17
Brand name 3	a1-a8，B13，B16

III. CUSTOMER MINING MODEL OF DIFFERENT BRANDS OF ELECTRIC VEHICLES

A. Overall idea of the model

In fact, the customer mining model to be established in this section is a binary classification model for whether to buy or not, so we consider the SVM (Support Vector Machine) algorithm. The basic model of the SVM is to find the best separation hyperplane on the feature space with the largest positive and negative sample intervals on the training set. Considering that this problem requires machine learning processing, this article changes the purchase intention to be 0 to-1, -1 means no purchase, and 1 means purchase. In addition, because it is a binary classification problem, and its small classification dimension and small data volume, we consider using the SVM algorithm to establish a shopping willingness prediction model and solve it. The advantage of the SVM algorithm is that it can handle classification and even multiple classification problems well, but because the SVC (Support Vector Classification) used in this paper is a time complexity O(n^2) algorithm, its operation speed is greatly reduced for high dimensions (e. g., 1000 dimensions) and high data volume (e. g., 1 million bars).However, it is calculated that the data selected in this paper meet the computable range of SVC on an ordinary computer, so this model is used.SVC is a support vector machine for classification, so this article is not going to go into much detail.

B. Model establishment

As for data segmentation, the first thing that comes to mind in this paper is the perceptron. The principle is to find a straight line and separate the data. If it rises to the high latitude, it is found that a hyperplane separates the data at the high latitude. The number of hyperplane of the perceptron may be at infinity, so the support vector machine model can be understood as finding the best one[5].

a) Functional interval

Function interval is the expression of the loss function of the perceptron, and can be expressed as:

$$\gamma' = \mathrm{y}(\omega^{\mathrm{T}}\mathrm{x}+\mathrm{b}) \quad (2)$$

b) Geometric interval

Geometric interval is the true distance from the point to the hyperplane, namely the function interval divides the normal vector:

$$\gamma = \frac{\mathrm{y}(\omega^{\mathrm{T}}\mathrm{x}+\mathrm{b})}{||\omega||_2} = \frac{\gamma'}{||\omega||_2} \quad (3)$$

As shown in Figure 2, the classification hyperplane $\omega^{\mathrm{T}}\mathrm{x}+\mathrm{b}=0$ can not only separate all samples, but also maintain a certain functional distance from the nearest sample point (support vector) (this function distance is 1), then such a classification hyperplane is better than the classification hyperplane of the perceptron. It can be shown that there is only one such hyperplane[6].

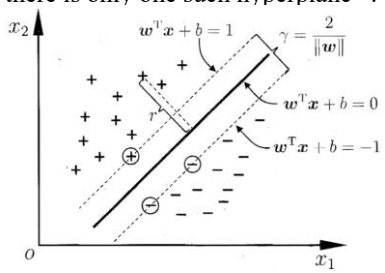


Figure 2. Support vector and interval

The SVM algorithm is to make all distances to the hyperplane greater than a certain distance, that is, all classification points are to be on both sides of the support vector of the respective categories. The mathematical formula is expressed as:

$$\max \gamma = \frac{\mathrm{y}(\omega^{\mathrm{T}}\mathrm{x}+\mathrm{b})}{||\omega||_2} \ \mathrm{s.t}\ y_i(\omega^{\mathrm{T}}x_i+\mathrm{b}) = \gamma'^{(\mathrm{i})} \geq \gamma'(\mathrm{i}=1,2,\ldots,\mathrm{m}) \quad (4)$$

Generally, this paper takes the function interval $\mathrm{y}'=1$, so that the optimization function of this paper is defined as:

$$\max \frac{1}{||\omega||_2} \ \mathrm{s.t}\ y_i(\omega^{\mathrm{T}}x_i+\mathrm{b}) \geq 1(\mathrm{i}=1,2,\ldots,\mathrm{m}) \quad (5)$$

The upper equation is equal to:

$$\min \frac{1}{2}||\omega||_2 \ \mathrm{s.t}\ y_i(\omega^{\mathrm{T}}x_i+\mathrm{b}) \geq 1(\mathrm{i}=1,2,\ldots,\mathrm{m}) \quad (6)$$

C. Oversampling of the samples

In the selected sample data, the number of negative samples far exceeds the number of positive samples (954: 20), so you cannot use the random sample segmentation of svc_train_test_split for learning samples directly. In random cases, the samples marked as supported may not be selected in the training set. In view of this situation, the SMOTE method provided by sklearn is adopted in this paper. For this serious imbalance, predictions are often biased that classification results are biased towards more observed classes. The easiest way to deal with this problem is to construct 1: 1 data and then remove some of the more categories (i. e., undersampled), or the less categories are Bootstrap sampled (i. e., oversampled). It's a problem. Firstly, the excluded data will lead to the loss of some hidden information. In the second, a simple

replication of the put-back sampling makes the model overfitting.

The basic idea of the SMOTE algorithm is to analyze and simulate a few categories of samples and add new samples simulated manually to the dataset, thus making the categories in the original data not seriously out of balance. The simulation process of the algorithm adopts k-Nearest Neighbors technology, simulated to generate new samples, and artificially adds partial noise to entry, making the data more credible.

Similarly, because of the sample imbalance, after using SMOTE, the data not oversampled can still be weighted with a higher weight, further increasing the training of the model. The dimension reduction diagram of each brand is shown in Figure 3.

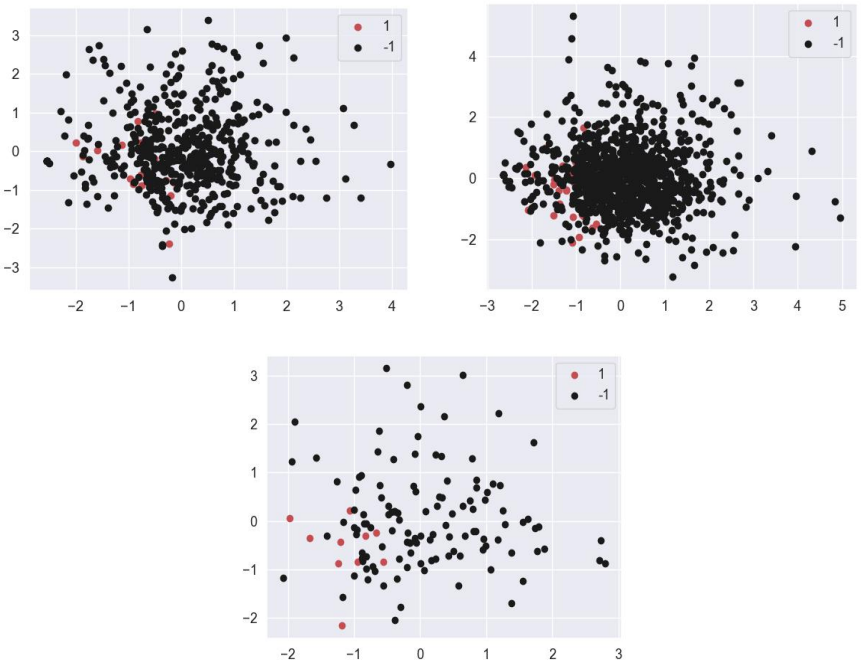


Figure 3. Pca(Principal Component Analysis) dimension reduction diagram of brand 1 (upper left), brand 2 (upper right) and brand 3 (under)

```
[[32  0]
 [ 4 36]]
report:
              precision    recall  f1-score   support

          -1       0.89      1.00      0.94        32
           1       1.00      0.90      0.95        40

    accuracy                           0.94        72
   macro avg       0.94      0.95      0.94        72
weighted avg       0.95      0.94      0.94        72

train_data: 1.0
test_data: 0.9444444444444444
```

```
[[356   4]
 [  2 355]]
report:
              precision    recall  f1-score   support

          -1       0.99      0.99      0.99       360
           1       0.99      0.99      0.99       357

    accuracy                           0.99       717
   macro avg       0.99      0.99      0.99       717
weighted avg       0.99      0.99      0.99       717

train_data: 1.0
test_data: 0.9916317991631799
```

```
[[32  0]
 [ 1 39]]
report:
              precision    recall  f1-score   support

          -1       0.97      1.00      0.98        32
           1       1.00      0.97      0.99        40

    accuracy                           0.99        72
   macro avg       0.98      0.99      0.99        72
weighted avg       0.99      0.99      0.99        72

train_data: 1.0
test_data: 0.9861111111111112
```

Figure 4. Adjustment parameters of brand 1 (upper left), brand 2 (upper right) and brand 3 (under)

D. SVM parameter tuning

We are using support vector machine to complete the entire model building so the parameters of svc should be

made certain adjustments to achieve the best results. The adjustment parameters of each brand are shown in Figure 4.

- Kernel function: the kernel functions used in this paper is chosen between the linear and radial basis function Gauss. Linear kernel: mainly used for linear separable cases. Less parameters, faster speed, the classification effect for general data is ideal. RBF kernel: mainly used for linear non separable cases. There are many parameters, and the classification results depend on them. Many people cross-verify the training data to find the appropriate parameters, but this process is time consuming. Because of the small number of samples, the RBF kernel function used has also achieved good verification results[7-8].
- Sample segmentation: 70% samples were learned and 30% samples were tested.
- Classification mode: because it is a binary classification problem, so the 'One-vs-One' mode is used for learning, which is the one-to-one classification method.
- Penalty parameter: the penalty parameter is processed for linear not separable data points. Since the data in this paper is not separable linear, the general penalty parameter value 10 is used to correct the data beyond the hyperplane.

IV. TEST THE POSSIBILITY OF 15 TARGET CUSTOMERS BUYING ELECTRIC VEHICLES

When using support vector machine to solve, it is found that SVC.fit (X, y) cannot achieve ideal results, and its data is not randomly separated, so the test set is recognized by the machine when it is implemented. After some modifications, we amended it to svm.SVC, so that the size and random range of the test set and the training set can be specified. However, in the cross-test, this paper found that the reliability of the data is not high, which leads to the following results: negative attitude is valid, but positive attitude is not. Therefore, analysis of its data set shows that the data has great limitations, with less supporting data and more supporting opposing data. Therefore, the way that decided to use oversampling is to artificially add the noise-containing data, and then balance the raw data in a weighted way, so that the raw data has a higher weight and the manually generated data has a lower confidence. After integrating the data through the above processing, the model has enough data for training. Finally, the predict_proba function is used to judge the supporting level of the model for each data. Then we can judge the new data set. The results are shown in Table 2:

TABLE II. PURCHASE USER FORECAST

Brand name	Purchase user number
Brand name 1	2，4
Brand name 2	7
Brand name 3	12

The prediction model is based on SVM algorithm, those support vectors with good discrimination ability for classification can be automatically found by learning the algorithm. The constructed classifier can maximize the interval between class and class, thus have better adaptation ability and higher classification rate, and the method only needs the category of boundary samples of various domains to determine the final classification results. After the above operation, the data are cross-verified by k-fold. The obtained values mainly observe the results of f1−socre, and the closer to 1, the better the model can judge the label. Thus this comprehensive prediction accuracy of this model is as high as 0.94.

V. CONCLUDING REMARKS

Firstly, the SVM algorithm in this paper can greatly simplify the classification problem. Secondly, its final decision function is determined by only a few support vectors, and the computational complexity depends on the number of support vectors rather than the dimension of the sample space, avoiding the "dimensionality disaster". In addition, a few support vectors determine the final result, which can not only help us capture the key samples and "eliminate" a large number of redundant samples, but also make the method not only algorithmically simple, but also "robust". Follow-up research can be completed by setting the weights of dimension. In each multidimensional dimension trained, each dimension weight is uniform, which against data with unique characteristics. These data can be analyzed, weighted by each vector and then added to learning.

REFERENCES

[1] Chen lei, Wang peiyong. prediction model of primary water supply pipe leakage time based on genetic least squares support vector machine [J]. journal of Zhejiang university of technology, 2021,(05):546-549.

[2] Yang Di, Fang Yangxin, Zhou Yan. Research on new classification based on MEB and SVM [J]. Journal of Guangxi Normal University (Natural Science Edition), 2021,(05):1-10.

[3] Hu Xuan, Li Chun, Ye Kehua. Application of support vector machine optimized by grey wolf algorithm in fault diagnosis of wind turbine gearbox [J]. Mechanical Strength, 2021,(05):1026-1034.

[4] Liu Chenyang, Xu Huang Rong, Duan Feng, Wang Taisheng, Lu Zhenwu, Yu Weixing. Spectral identification of rabbit liver VX2 tumor based on genetic algorithm and support vector machine [J]. Spectroscopy and Spectral Analysis, 2021,(10):3123-3128.

[5] Astuti Suryani Dyah,Tamimi Mohammad H.,Pradhana Anak A.S.,Alamsyah Kartika A.,Purnobasuki Hery,Khasanah Miratul,Susilo Yunus,Triyana Kuwat, Kashif Muhammad, Syahrom Ardiyansyah.Gas sensor array to classify the chicken meat with E. coli contaminant by using random forest and support vector machine[J].Biosensors and Bioelectronics: X,2021,9(9):5-8.

[6] Tan Hongchuang,Xie Suchao,Liu Runda, Ma Wen.Bearing fault identification based on stacking modified composite multiscale dispersion entropy and optimised support vector machine[J].Measurement,2021,186(186):6-.

[7] Xiao Shijie, Wang Qiaohua, Li Chunfang, Zhao Limei, Liu Xinya, Lu Shiyu, Zhang Shujun. Construction of milk purchasing grading model based on random leapfrog and support vector machine [J]. Smart Agriculture (English and Chinese), 2021,(05):1-9.

[8] Li Mengmeng, Liu Jingdang, Liang Tianyi, Tan Liang, Wang Gang, Zhu Xi. Prediction model of sulfur element in magmatic sulfide deposits based on support vector machine algorithm [J]. Journal of Jilin University (Earth Science Edition), 2021,(05):1-15.

A study of the influence of different opinion leaders on the evolution of followers' views in social e-commerce networks

Jianhong Jiang
School of Business
Guilin University of Electronic Technology
Guilin China
e-mail: jjhome@guet.edu.cn

Li Wang
School of Business
Guilin University of Electronic Technology
Guilin China
e-mail: wangli1148@126.com

***Abstract*—In the recently emerged social e-commerce platforms, online opinion leaders play a crucial role. Based on the finite trust model in opinion dynamics and combined with the actual situation of social e-commerce platforms and research literature, this paper divides opinion leaders into 'Big-V' type opinion leaders and authoritative opinion leaders, and divides followers into ordinary followers and fans, and constructs different opinion leaders in social e-commerce networks based on bounded trust and self-confidence degrees - the follower opinion evolution model. The influence of different opinion leaders on the evolution of followers' views is explored through multi-agent simulation. It is found that opinion leaders play an important role in the aggregation of views, and the breadth of follower coverage of 'Big-V' type opinion leaders makes them more influential and able to gain more strong supporters than authority-type opinion leaders.**

Keywords-opinion dynamics theory,social e-commerce, opinion leaders,multi-agent emulation

I. INTRODUCTION

In the current e-commerce industry, users mostly need to check the historical reviews about the product before purchasing it, but the false recommendations and massive redundant recommendations seriously affects the effectiveness and accuracy of users' access to recommendations [1]. Nowadays, with the integration and development of e-commerce and social networks, professional social networking platforms for product sharing, evaluation and communication have started to emerge and become important channels for users to exchange views and proliferate products (e.g. Weibo, Xiaohongshu, Facebook, etc.). In these sites, users can establish online social relationships to share product information or reviews, and can also ask product-related questions to network members. Compared to traditional e-commerce platforms, the establishment of perceived trust by consumers here relies more on recommendations from others in the form of user-generated content [2]. In this process, we can find individuals who can exert influence on the opinions, decisions and actions of most other consumers, i.e., online opinion leaders, who play a crucial role in social e-commerce networks. Therefore research on the influence of online opinion leaders on groups has also received extensive attention. Qing et al [3] studied the influence of online opinion leaders' mobilization methods on Internet users' online cluster behavior participation in the context of product harm crisis. Liu Li et al [4] used an empirical research method to test the influence of opinion leaders on college students' participation in tourism virtual community behavior by constructing a conceptual model. Tobon[5] looks at the role of opinion leader's electronic word of mouth as well as the effect of its valence and product type on the decision to buy or not buy from a realistic experimental online store design.

In social e-commerce platforms, the relationships between individuals interacting and influencing each other constitute a social network. In recent years, the development of statistical physics, multi-agent models and complex networks has provided researchers with tools for quantitative analysis or numerical simulation of large-scale social networks. Among them, dynamic simulations of perspectives using opinion dynamics models can better reflect the process and mechanism of perspective evolution [6]. At the same time, in order to more detail the change of individual opinions in social networks, more and more scholars focus on microscopic opinion dynamics models based on multi-agent system. Tingyu Li et al [7] constructed an opinion dynamics model with the addition of media influence variables to study the key role of media in the evolution of opinions; Giacomo[8] consider the modeling of opinion dynamics over time dependent large scale networks.

In this paper, we take social e-commerce platform as a background, construct a social network in line with its characteristics, set up different types of opinion leaders and followers based on literature studies and realistic situations. Then we introduce a bounded trust model of opinion dynamics (Hegselmann-Krause model) and improve it. Through multi-agent simulation, we further explore the impact of different types of opinions leaders on followers in social e-commerce backgrounds. A theoretical basis is provided for companies to design appropriate marketing tools and opinion guidance for their brands and products, and feasible measures are proposed.

II. THE CONSTRUCTION OF A SOCIAL E-COMMERCE NETWORK

The metrics of mainstream social networks from Stanford University's publicly available social network dataset show that mainstream social network, such as Facebook, Twitter, etc, have both small-world properties and scale-free properties. In order to evolve the model on a network structure similar to real social networks, we need to select a network with both scale-free and small clustering properties, referring to the network construction approach of Gong Yi Li [9], the process of constructing the network, the generative edge algorithm fuses preferential connections and neighborhood connections. Preferential connection means that new node will be prioritized to

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00019

establish a connection with a large influence in the network; the neighborhood connection refers to the node, which is preferentially generated from the node from the near. The improved network model generate edges with a preferential connection with a certain probability to the new join node (set this probability of p) and according to the probability of 1-p in the neighborhood connection way. Assuming p = 0.7 for preferential connectivity and 1-p = 0.3 for neighborhood connectivity, the model evolved to a number of nodes of 1000 is shown in Fig.1. The average clustering coefficient of the improved network is 0.5021. From Stanford University public SNAP data sources, the average cluster coefficient of Twitter social network is 0.5653[10]. It can be seen that the improvement results are in line with the actual situation.

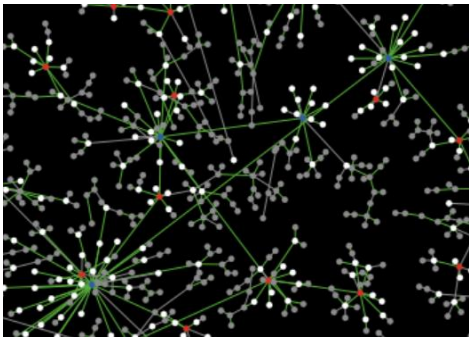


Figure 1. Simulation of the evolution of real social networks

III. MODELING

A. *Opinion leader and follower settings*

The traditional literature divides crowds into two broad categories: opinion leaders and followers, they differ in their adherence to their views, and the views of opinion leaders can influence the views of followers [11]. Zhu [12] considers the difference in the acceptance and communication of the viewpoint, it will be refined as a fan and ordinary follower, and the filaments are subdivided into a fanatic hot fans and ordinary fans. However, few studies have considered the diversity of opinion leaders and the differences in their influence. Therefore, this paper considers the characteristics of social e-commerce networks and the diversity of opinion leaders. We refer to Shao Peng [13] to define three special types of users who influence the evolution of group views, and subdivides opinion leaders into authoritative opinion leaders and 'Big -V' type opinion leaders (Hereinafter referred to as V opinion leader). Authoritative opinion leaders have a high degree of expertise in a certain field and are able to provide objective and authoritative advice to consumers on social e-commerce platforms, such as scholars, experts. And by posting information with expertise over time, these users have built up a high level of trust in the network. In turn, they themselves have a smaller trust threshold and do not easily change and update their opinions. V opinion leaders are those who have a strong influence and a large number of social links in social commerce platforms, which are crucial for product adoption and spreading. They generally harvest followers and influence their purchase decisions through emotional, subjective and interesting information postings. They themselves have a low trust threshold and are influenced by other opinion leaders.

The followers in the social e-commerce network are set as general followers as well as fans. The number of fanatic fans on this platform is small as most of the consumers on this platform seek sharing and advice on product purchase, so it is not set in this paper. Fans in this model are followers of opinion leaders who are more easily persuaded by opinion leaders, usually show super interest and enthusiasm for the product, and are more active in product proliferation. Ordinary followers, on the other hand, are consumers who do not follow opinion leaders, who are more easily persuaded by those around them, and show no interest in product diffusion.

According to the above definition and previous studies on the structure of social network opinion leaders [14], "medium V" users have stronger opinion guiding power and stronger professionalism, which is equivalent to the authoritative opinion leaders in this paper, while "medium V" users have fewer followers and lower influence compared with "big V" users. Therefore, in the network where opinion leaders exist simultaneously, the top 1% of nodes with the highest degree value are set as V opinion leaders, and the nodes with degree values between the top 2% and 5% are set as authoritative opinion leaders.

B. *A model for the evolution of different opinion leader-follower views based on bounded trust and confidence levels*

In this paper, we choose to introduce the HK model of "confidence factor". The "confidence factor" is the individual's self-confidence and belief that the vast majority of people do not completely lose their views in a clash of views, but retain them to some extent.

Set up the network with N individuals, among which there are N_0 opinion leaders, N_1 regular followers, and N_2 fans. Among the N_0 opinion leaders, the number of authoritative opinion leaders is Y_1 and the number of V opinion leaders is Y_2. When the opinions at time t of all individuals are denoted by $x_i(t)$, for i = 1,…,N. The completely positive and completely negative opinions are, respectively, defined by $x_i(t)$ = 1 or -1. Assume that the initial viewpoint X(0) of all individuals conforms to a uniform distribution of [-1, 1].

If a fan has a total number of neighbors j of m, with the number of followers m_1, the number of authoritative opinion leaders m_2 and the number of V opinion leaders m_3, then its opinion update rule is:

$$x_i(t+1)=\alpha_i \frac{1}{m_i^A}\sum_{j=m_1+1}^{m_1+m_2} a_{ij}(t)x_j(t)+\beta_i \frac{1}{m_i^V}\sum_{j=m_1+m_2+1}^{m_1+m_2+m_3} a_{ij}(t)x_j(t)$$

$$+\left(1-\alpha_i-\beta_i-\gamma_i\right)\frac{1}{m_i^F}\sum_{j=1}^{m_1} a_{ij}(t)x_j(t)+\gamma_i x_i(t) \quad (1)$$

where, i=1,2,……N_2, $a_{ij}=\begin{cases}1,\text{if } |x_i(t)-x_j(t)|\le\varepsilon_i\\ 0,\text{if } |x_i(t)-x_j(t)|>\varepsilon_i\end{cases}$

$m_i^A=\sum_{j=m_1+1}^{m_1+m_2} a_{ij}(t)$, $m_i^V=\sum_{j=m_1+m_2+1}^{m_1+m_2+m_3} a_{ij}(t)$, $m_i^F=\sum_{j=1}^{m_1} a_{ij}(t)$ denote the number of authoritative opinion leaders, V opinion leaders, and followers in individual i's neighborhood that are within the confidence level, respectively. Coefficients α_i,β_i denote the confidence

level of individual i in authoritative opinion leaders and V opinion leaders, respectively. Due to the different traits of the two, setting$\alpha_i>\beta_i$, indicates that individual fans trust authoritative opinion leaders more. Also set $m_2 < m_3$, since V opinion leaders have more followers and wider influence, then correspondingly fans can receive more messages from V opinion leaders. Coefficient γ_i indicates the confidence coefficient of an individual. The larger it is the less it is influenced by the outside world and the more it trusts its own judgment.

The opinion updates of ordinary followers are less influenced by opinion leaders than ordinary fans. Assuming that the total number of neighbors j is m, where m consists of fans and regular followers whose opinion update rule is:

$$x_i(t+1)=(1-\gamma_i)\frac{1}{m_i^F}\sum_{j=1}^{m}a_{ij}(t)x_j(t)+\gamma_i x_i(t) \quad (2)$$

where i=1,2,……N_2, the means of a_{ij}、γ_i are consistent with the above.

The views of authoritative opinion leaders are not influenced by other opinion leaders and followers, but are derived from their own knowledge and experience and from research findings that are more authoritative than their own, and the rules for updating their views are:

$$x_i(t+1)=\gamma_i x_i(t) \quad (3)$$

where i=1,2, … … Y_2, $\gamma_i \in [0.8,1]$ indicates that authoritative opinion leaders will have to revise their views appropriately based on the latest professional information.

The opinion of a V opinion leader is not influenced by followers, but will be influenced by opinion leaders in the neighborhood. Suppose the total number of neighbors of j is m, among which those who are followers are m_1, those who are authoritative opinion leaders are m_2, and those who are V opinion leaders are m_3 whose opinion update rule is:

$$x_i(t+1)=\alpha_i\frac{1}{m_i^A}\sum_{j=m_1+1}^{m_1+m_2}a_{ij}(t)x_j(t)+\beta_i\frac{1}{m_i^V}\sum_{j=m_1+m_2+1}^{m_1+m_2+m_3}a_{ij}(t)x_j(t)+\gamma_i x_i(t) \quad (4)$$

where, i=1,2,……Y_2, the means of a_{ij}、β_i are consistent with the above.

In this model, V opinion leaders take their confidence level to a consistent level of 0.25. According to the literature, there is no confidence level for authoritative opinion leaders, and the confidence level for followers fits a uniform distribution of [0,1] and hold the individual opinion values constant when all individual neighbors are outside their confidence levels. Denote the set of neighbors $N_i(t)$ that are within the i-confidence level, i.e., $x_i(t + 1) = x_i(t)$ when $N_i(t)$ is null.

IV. SIMULATION AND PARAMETER SETTING

Using netlogo software, we construct a network and add the above-mentioned opinion update rules and set the output of the line graph. Based on previous research, this paper investigates the evolution of public opinion when there are two competing products and the opinion leaders hold different opinions, thus studying the influence of different types of opinion leaders.

A. Evolution of views when different types of opinion leaders support different competing products

Set the opinion of the V opinion leader to 1, the opinion of the authoritative opinion leader to -1, and set the confidence of the opinion leader γ =1, confidence level ε = 1 to indicate that different types of opinion leaders firmly support different competing products respectively, and the opinion leaders' views do not change during the evolution. The evolution of the views of fans and ordinary followers is shown in Fig 2, Fig 3.

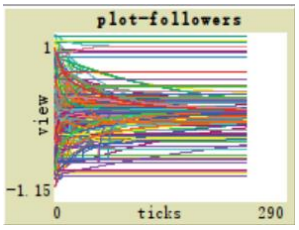


Figure 2

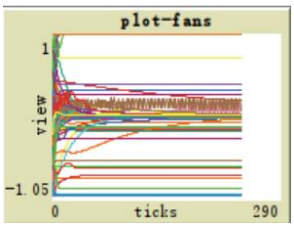


Figure 3

From Fig 2 and Fig 3, we can see that both ordinary followers and fans are influenced by opinion leaders. Fans' viewpoint polymerization effect is better, ordinary followers' viewpoint more dispersed. Most followers tend to be close to neutral views, while a small number of views are maintained by extreme individuals. In addition, the influence of opinion leaders can spread from fans to ordinary followers. By comparing the number of fans with extreme views in the fan evolution chart, we can see that the number of opinion supporters of V opinion leaders is higher after the evolution.

B. Evolution of views when a certain type of opinion leader supports a competing product

This part of the simulation sets the opinion value of V opinion leaders to 1, and the opinion value of authoritative opinion leaders obeys uniform distribution and tends to be neutral, that is, the self-confidence of authoritative opinion leaders is set γ =0.8. This means that the confidence level of the authoritative opinion leader is set to 1, which means that the V opinion leader supports a product, while the authoritative opinion leader is in a wait-and-see state. The evolution of the opinions of fans and ordinary followers is shown in Fig 4 and Fig 5.

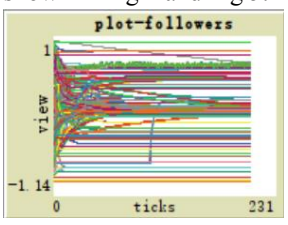


Figure 4

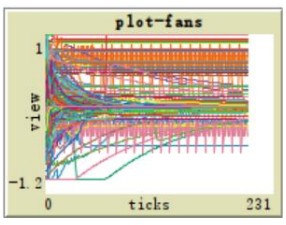


Figure 5

From the figure, we can see that followers are roughly divided into two opinions, neutral and positive point views, but there will be up and down floats. Comparing with the simulation diagram in Part I, when opinion leaders are close enough to their fans' views or have been in contact with them for long enough, they can strongly influence

them; however, it is not yet possible to see which type of opinion leader is at work.

In order to better study the influence of the two opinion leaders, the opinion value of the authoritative opinion leader is set to 1, and the opinion value of the V opinion leader follows a uniform distribution and tends to be neutral, i.e., the confidence of the V opinion leader is set to γ =0.8. This indicates that the authoritative opinion leaders support a competing product, while the V opinion leaders are in a wait-and-see situation. The evolution of the opinions of fans and ordinary followers is shown in Fig 6 and Fig 7.

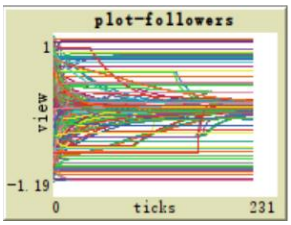


Figure 6

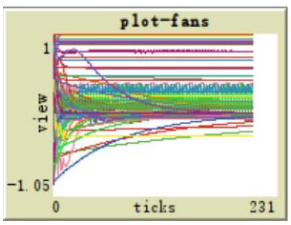


Figure 7

The main difference between the evolution diagram of this case and the previous case is that there are no two more distinct clusters of opinions, only in neutral view attachment aggregation. This comparison shows that V opinion leaders are more influential than authoritative opinion leaders in attracting followers with strong support tendencies or extreme views. Combined with the e-commerce scenario, if a company attracts strongly supportive followers accordingly, it may first consider enlisting the support of V opinion leaders.

V. CONCLUSIONS AND MANAGERIAL IMPLICATIONS

This paper extends the HK model in continuous opinion dynamics by combining today's popular social e-commerce scenarios, constructs a social e-commerce network with both small-world and scale-free properties, and combines the classification types of opinion leaders and follower with the HK model to form different opinion leader-follower viewpoint evolution models based on bounded trust and assertiveness. In addition, integrating e-commerce into opinion dynamics extends the study of opinion dynamics by examining the presence of rival contenders in the network, and the situation of contender supporters. It is found that when two opinion leaders have opposing views on the platform, most of the individual views will be guided to neutral views and some strong supporters will be lost, which is detrimental to the development of competing products; in the case where one side supports the view and the other side holds a neutral view, the V opinion leader with a strong supportive view is more able to attract followers with strong supportive tendencies or extreme views, making the group divided into two parts with different opinions; when the authoritative opinion leader holds a strongly supportive view, the aggregation of group views is not effective, and most individuals tend to have a neutral view.

In real social ecommerce scenarios, opinion leaders can powerfully influence followers when their opinions are close enough to those of their followers, or when they have been in contact for long enough. When there are opposing competing products, in order to retain and gain more strongly supportive followers one should actively communicate with opinion leaders with opposing views to try to make them eliminate their bad opinions about the product and give priority to gaining the support of V opinion leaders, whose breadth of fan coverage makes them more influential. Exploring the sources of influence of different opinion leaders and constructing simulation networks from real social network data can be considered in future studies.

ACKNOWLEDGMENT

This work was financially supported by the National Natural Science Foundation of China project (71940008), Ministry of Education Humanities Social Sciences Youth Project(17YJCZH074) and Guilin University of Electronic Science and Technology Graduate Education Innovation Project(2019YCXS069),

REFERENCES

[1] Yu Zhen, Zhu Jie, Shen Guicheng. An electronic commerce trust model and simulation based on social network [J]. System Simulation, 2018, 30 (08): 3115-3123.

[2] Lount Robert B. The impact of positive mood on trust in interpersonal and intergroup interactions.[J]. Journal of personality and social psychology,2010,98(3):

[3] Qingping, Zhang Ying, Tu Ming, Zhang Yong.Affect of Network Opinions Leaders Mobilization Mobilization on Network Cluster Behavior - Based on Experimental Research Based on Product Injury Crisis [J]. Management World, 2016 (07) : 109-120.

[4] Liu Li, Cheng Dong, Su Xin. Empirical Study on the Influence of Opinion Leaders' Participation in Tourism Virtual Community - A Case Study of College Students [J] .Journal of Tourism, 2018,33 (09): 83-93.

[5] Tobon Sandra,GarcíaMadariaga Jesús. The Influence of Opinion Leaders' eWOM on Online Consumer Decisions: A Study on Social Influence[J]. Journal of Theoretical and Applied Electronic Commerce Research,2021,16(4):

[6] KRAUSE U.Positive nonlinear difference equations [J] .Nonlinear Analysis Theory Methods & Applications, 1997,30.

[7] Tingyu Li,Hengmin Zhu. Effect of the media on the opinion dynamics in online social networks[J]. Physica A: Statistical Mechanics and its Applications,2020,551:.

[8] Giacomo Albi,Lorenzo Pareschi,Mattia Zanella. Opinion dynamics over complex networks: Kinetic modelling and numerical methods[J]. Kinetic & Related Models,2016,10(1):

[9] Li Gongyi. A Study on the Evolutionary Model of Opinions Leaders and Memory Ways [D]. Huazhong University of Science and Technology, 2019.

[10] VillalbaDiez Javier,Molina Martin,Schmidt Daniel. Geometric Deep Lean Learning: Evaluation Using a Twitter Social Network[J]. Applied Sciences,2021,11(15):

[11] Viju Raghupathi,Joshua Fogel. Facebook advertisements and purchase of weight-loss products[J]. Journal of Medical Marketing: Device, Diagnostic and Pharmaceutical Marketing,2013,13(4):

[12] Zhu Xiaoxia, Meng Jianfang. Study on the Evolution of Opinion Leaders and Followers in Multi - level Network - Based on Acritical Trust Model [J] Intelligence Science, 2019,37 (06): 30-36.

[13] Shao Peng, Hu Ping.Effects of special users of complex networks on population viewpoint evolution [J] .Journal of Electronic Science and Technology, 2019,48 (04): 604-612.

[14] Cao Yu, Zhang Zhan. Research on Weibo Opinion Leaders Based on Media Power Structure [J] .Print, 2016 (09): 43-49.

Level-based E-government Cloud Cross-domain Access Control Technology

Jing Gao
Jilin Institute of Chemical Technology
Jilin, China
e-mail: 836076585@qq.com

***Abstract*—The application of the cloud computing in the field of e-government is an important aspect of China's informatization construction. The e-government cloud can increase the efficiency of government offices and promote the development of a harmonious society. The current e-government system must meet the requirements of hierarchical protection standards. How to seamlessly migrate the user management of the existing e-government system to the e-government cloud system and achieve cross-domain secure access based on the user level is a hot and difficult issue for the e-government cloud. Therefore, this paper proposes the novel level-based E-government cloud cross-domain access control technology. The proposed model will guarantee the safety of the model compared with the other state-of-the-art methodologies. The simulation results prove the performance.**

Keywords-E-government cloud;hierarchy;cross-domain access control; access control management

I. INTRODUCTION

Since the launch of Internet technology, it has been in a state of rapid development. With the continuous creation of new achievements, Internet technology has been closely linked with people's lives and has become a powerful force that promotes society, economy, and even politics, and it is also one of the countries in the world that displays comprehensive national strength. As one of the representative countries, China has already played an important role in the international arena. At this stage, China is continuously developing Internet technologies and has made some achievements in exploring and advancing. Internet technology has been integrated into all aspects of society. Among them, the field of e-government is also an aspect of our country's attention, such as the use of e-government to serve the people more efficiently, improve government functions, and promote the development of a harmonious society.

E-government is a significant trend in social development. Countries all over the world are developing e-government to promote the comprehensive and convenient government services, as well as the stable development of society. China's Internet technology is in a state of rapid development, but it is relatively late to start, so China's e-government is still in the initial stage of development. Fortunately, academic research, social practice, and national investment in Internet technology, have all contributed to the development of e-government.

At the same time, in the process of development, it also faces some problems, such as how to develop e-government more comprehensively, make the scale of e-government larger that provide the more comprehensive services, and provide fine-grained services to the people. Therefore, in the process of promoting e-government, we must not only pay attention to speed, but also pay attention to fineness.

Cloud computing is a new chapter of Internet technology at the present stage. It is a representative of the new scientific and technological forces at the present stage. It has originated from abroad and has created an upsurge in the Internet industry. It is the future direction of the IT industry and affects electronics in its unique way. The development of government affairs, cloud computing can become a new opportunity for Internet technology at this stage. E-government supported by Internet technology must seize this opportunity to develop rapidly, establish a continuous e-government cloud system, and achieve efficient administrative system management to provide more convenient services and more comprehensive service.

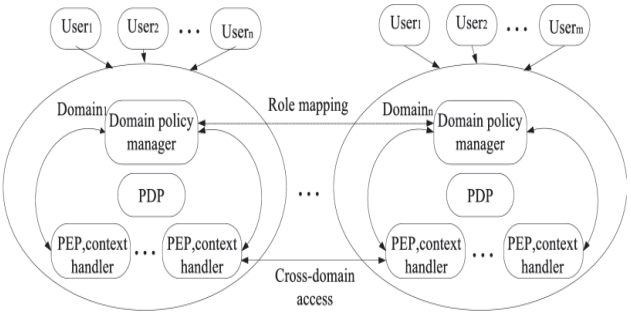


Figure 1. The Access Control with Multiple Users

The e-government system must meet the requirements of hierarchical protection standards, how to seamlessly migrate the user management of the existing e-government system to the e-government cloud system, and achieve cross-domain security access based on the user level, which is a hot issue for the e-government cloud. This paper examines how to implement cross-domain access to the e-government cloud based on user ratings, and explores security issues that solve the problem of the inadequate granularity of access control when users share resources. There is a serious problem of low resource integration in e-government. All departments at the county level have their own information platform, independently releasing information, with low correlation degree, weak information interactivity and serious information asymmetry.

At present, the government holds the social resources in the most valuable information and information database, due to the lack of the information flow, transfer and diffusion, useless information, distortion of information spread on the network that lead to enterprises and

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00020

individuals can't get all the required information and the services through formal channels. According to the characteristics of the electronic government affairs information service, comprehensive SERVQUAL model and the scholars put forward the network information service quality evaluation model, the process of using e-government service from the public's perceptions of the quality of modern information service, in the e-government information service of public satisfaction model is established.

II. Silent Aesthetics' Inspiration to Chinese Aesthetics

A. Cross-domain Access Control Technology based on Cloud Computing

While cloud computing exerts its powerful advantages, it also accompanies the user information leakage, data destruction, loss, and stealing in the cloud system. When cloud computing is applied to e-government, users in different security domains need to share resources and need higher security protection. The typical solution is identity and access control management.

IAM is one of the important products of cloud computing development. RBAC occupies an important position in the research and development of IAM. Although RBAC is a traditional access control technology, its improvement schemes are endless. The current domestic development situation is lower than that of foreign countries, but domestic researchers are also actively exploring.

The main technical support for identity federation is SAML. Typical use cases include cross-domain single sign-on, user account configuration, rights management, and user attributes.

The access control mechanism originated in the 1970s and aims to protect resources and ensure that resources are legally accessed. The main participants in the access control include three links: subject, object, and access strategy. The subject must perform various access operations on the object according to the specified access rules. Of course, the operation must be within the authority of the subject. Access control allows only authorized users within the scope of rights to gain access to resources while also isolating illegal users and resources.

The main concepts of access control include subject, object, authority, and authorization. The entity refers to the entity that issues an access request, which is generally a system user or program process. The object is a passive entity that passively accepts the operation of other entities. Generally, it is a protected resource in the system, such as data and user information, infrastructure and so on. Permissions refer to the types of manipulations that can be performed by the subject in the access control.

B. Cross-domain Identity Management Mechanisms

Cloud Computing IAM is mainly implemented as a federation technology. Joining has become a key concept in identity management. It can not only reduce the complexity of identity management within the organization, but also provide better experience for users when accessing services. In this sense, the process of trust establishment allows fast and seamless interaction between different trust domains.

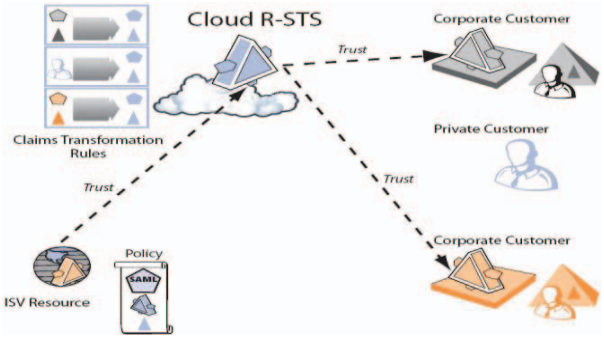


Figure 2. The Cloud R-STS Framework

Coalition has become a key concept of identity management. Its main goal is to apply certain policy transmissions and share user attributes in different security domains. The federated model allows users in one domain to securely seamlessly access resources in another domain without the need for redundant user login processes. One of the most popular use cases is single sign-on (SSO), which allows users to obtain multiple sites by completing authentication on only one site. And do not provide redundant user attributes and to avoid multiple logins. Therefore, the task of identity management is separated from the entities that provide services, so as to reduce the complexity of user management, it is possible to focus on their core business and improve the user's interactive experience on various security platforms.

In this example, identity information is shared between mobile operators (IdP1) and travel agencies (IdP2), allowing users to log in only once and gain seamless access to services and applications in another security domain. The user's local domain IdP1 provides a local service. When the user enters the mapping service web page and provides his credentials (such as user name and password), he can obtain a link to the security domain IdP2 web page and access services or resources, such as accommodation or catering information. Identity federation has great advantages, including typical user cross-domain single sign-on, user account configuration, rights management, and user attribute management. The currently widely used joint technologies include SAML, OpenID, Liberty Alliance, WS-Federation, and Shibboleth.

C. Shibboleth Architecture

The Shibboleth architecture provides an identity management solution for resource sharing in a cross-domain scenario. This solution is an open source project. Currently 2.0 is the latest version of Shibboleth. Shibboleth uses SAML as its technical support to establish single sign-on between domains or within the domain. At the same time, secretive attribute information is passed in the authentication process, which protects the user's privacy to a certain extent.

The Shibboleth architecture is extended on the basis of SAML and involves identity providers (IDPs), service providers (SPs) and WAYFs (where are you from). The IDP mainly completes the creation, management, maintenance, and storage of user attribute information in

the domain, including authentication services, single sign-on services, resolution services, attribute services, directory services, and local authentication services. The SP is responsible for managing the resources in the domain. When the external domain user applies for access to the resources, it processes the user's request, makes a property request to the user IDP, and responds to the user according to the request result and access control policy, including the resource management service, the attribute request service, and Assertion invokes the service. WAYF is a bridge between IDP and SP, assisting in the completion of user interaction in the authentication of two domains.

Information security protection has always been the focus of China's information-related work. With the development of information technology, some related technologies have been accompanied by research. Hierarchical protection is an important safety guarantee technology. It mainly develops related access services based on host-guest level information and resource specifications.

The significance of hierarchical protection lies in the protection of confidential resources through the implementation of related hierarchical policies, and the assignment of systems, users, roles, resources, permissions, and other related components to hierarchical labels to specify access policies. The value of the in-depth study of this mechanism is to establish a more comprehensive specification for the secure sharing of resources, and at the same time to help achieve a more granular access control mechanism.

A complete hierarchical protection system must ensure the security of several components of the computing environment, network, management, and borders, and each component must also complete the corresponding rating. The computing environment is the component that completes a series of operations. The network is the component that ensures the information interaction between the computing environments. Management is responsible for providing a secure and reliable implementation infrastructure platform for the other three components. The boundary is defined for the system computing environment security zone and maintains the interaction, and interaction between the computing environment and the network.

D. Shibboleth Architecture

With the continuous improvement of the e-government system, the government has always attached great importance to the construction of the intranet. While using the existing network technology to advance the construction of the intranet of government affairs, the security requirements involved have also increased along with the continuous improvement of the system's functionality. At the same time, corresponding solutions have also been proposed for various risks that accompany them, such as hierarchical protection. Hierarchical protection ideas originate from the use of multi-level protection ideas by the U.S. military to save files. People and files are divided into four levels from high to low. However, only people with high file security can manipulate files, it has not been successfully applied to computer systems. Subsequently, we proposed that this hierarchical protection idea should be applied to the computer system and that the design of the computer system should be improved. Soon, two mathematicians, D.E.Bell and L.J.Lapadula, proposed the Bell-Lapadula (BLP) model, which proved from a mathematical point of view that the idea of hierarchical protection can be implemented on computers.

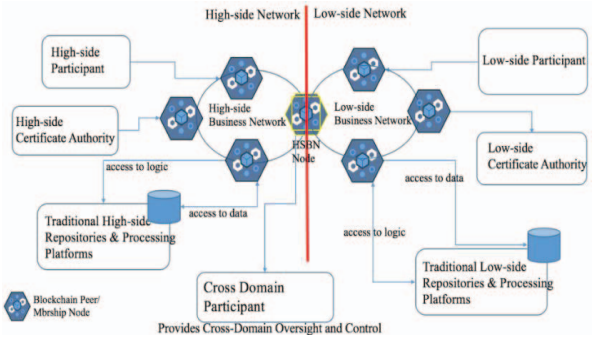


Figure 3. The Cloud Control Provider for References

The pattern of hierarchical protection ideas is closer to the distributed management of actual systems. If users, roles, and resources are ranked in a resource sharing cross-domain access system, fine-grained resource sharing can be achieved. Specifically, the intranet is divided into levels. The more levels are divided, the finer granularity of access control is. However, the cost is more complicated in user information management and maintenance. Therefore, the two are weighed according to the specific needs of the intranet. For example, after the classification, the result of the division may be a department. At this time, each department can become a domain, and each domain has its own level information. The core layer refers to the basic management function of the distributed resources, and provides distributed application deployment environment through abstract services. The core layer can implement abstract services through operating system kernels, hypervisors, virtual machines, or cluster middleware, providing a deployment environment for distributed application users. Therefore, this article in the analysis of the subject matter of the core relationship between the Foundation, proposed based on the clustering of object aggregation information level deduction method, the method by formal concept analysis, for the same security domain object for the similarity analysis, in order to achieve the object resource of the cluster, and based on the attribute or attribute subset level fuzzy sets possibility measures as deduced by a similar object to derive higher-level information of the possibility, and finally according to the aggregation information of the possibility level, develop appropriate access control policies to effectively reduce the multi-stage network leakage risks.

Cloud computing security standards can not only enable users to describe data security protection, and should support the user enterprise safety management requirements of users, in particular, such as the analysis of the log information to view, information collection, data use and investigation of illegal operation. At present, the cloud business model has not yet reached the degree of maturity, responsibility and authority between the users and the cloud computing service definition is not clear and

users and cloud computing services may be in the scope of control and limit conflicts. Information security rank protection system is dominated by national regulatory authorities. That is an important work of information system security. Along with step by step according to the steps and phases of the system implementation, supervision, inspection and guidance of the information system will become the future regulatory work content on a permanent basis. Regulatory object is important in the field of industry and information system, the safety status of regulatory focus on information system regulatory goal is based on a comprehensive grasp the condition of system security, information security policy, related to the nation decision-making to provide strong support. For the above reasons, it is necessary and urgent to build a national integrated monitoring system.

Cloud computing is undoubtedly one of the technical supports for the rapid development of e-government, and the promotion of e-government can provide cloud-advantaged services. However, e-government has benefited from cloud computing and has improved its functions while also facing corresponding security risks.

The metadata of IDP and SP must be synchronized with each other to ensure the smooth communication. Metadata is the data used to describe the attribute information, including name, attribute, type, etc. The information description of the entity is implemented in a standard format, and the IDP and the SP communicate, and it is necessary to confirm in the metadata whether the requesting party is a trusting relationship. Yes, you can proceed with the next communication, otherwise the communication will be terminated.

III. CONCLUSION

With the rapid development of the cloud computing, more and more participants contribute to the advancement of cloud computing. Some organizations or enterprises acquire existing cloud computing technologies and at the same time contribute new achievements and form a virtuous circle to promote faster development of cloud computing. An example of a typical development product is the OpenStack cloud platform, which now has a complete system that is deployed by organizations and enterprises to their own industries. This general solution realizes the cross-domain access control with user-level attributes in the e-government cloud system, and establishes IDP and SP systems, and achieves hierarchical protection of users and resources in the domain. In the future, we will apply the proposed model into more related scenarios.

REFERENCES

[1] Na, L., Yun-Wei, D., Tian-Wei, C., Chao, W., Yang, G. and Yu-Chen, Z., "Cross-Domain Authorization Management Model for Multi-Levels Hybrid Cloud Computing," International Journal of Security and Its Applications. India, 2015,vol. 9, pp. 357–366.

[2] Alhumrani, S.A. and Kar, J., "Cryptographic Protocols for Secure Cloud Computing," International Journal of Security and Its Applications. India,2016,vol. 10,pp.301-310.

[3] Alam, Q., Malik, S.U., Akhunzada, A., Choo, K.K.R., Tabbasum, S. and Alam, M., "A Cross Tenant Access Control (CTAC) model for cloud computing: formal specification and verification," IEEE Transactions on Information Forensics and Security. United States, 2017,vol. 12, pp.1259-1268.

[4] Zhou, Z., Gaaloul, W., Hung, P.C., Shu, L. and Tan, W., "IEEE access special session editorial: Big data services and computational intelligence for industrial systems," IEEE Access. United States, 2015,3, pp. 3085–3088.

[5] Shere, R., Shrivastava, S. and Pateriya, R.K., CloudSim Framework for Federation of identity management in Cloud Computing. 2017.

[6] Fritzsche, D., Grüninger, M., Baclawski, K., Bennett, M., Berg-Cross, G., Schneider, T., Sriram, R., Underwood, M. and Westerinen, A., "Ontology Summit 2016 Communique: Ontologies within semantic interoperability ecosystems," Applied Ontology. Netherlands, 2017, vol.12, pp.91-111.

[7] Cai, Z., Deng, L., Li, D., Yao, X., Cox, D. and Wang, H., "A FCM cluster:cloud networking model for intelligent transportation in the city of Macau," Cluster Computing. United States, 2017, 1-10.

[8] Xie, T., Li, C.D., Wei, Y.Y., Jiang, J.J. and Xie, R., "Cross-domain integrating and reasoning spaces for offsite nuclear emergency response," Safety science. Netherlands, 2016, vol.85, pp. 99–116.

[9] Liu, W., Liu, X., Liu, J. and Wu, Q., "Auditing Revocable Privacy-Preserving Access Control for EHRs in Clouds," The Computer Journal.England, 2017, vol.60, pp.1871-1888.

Trajectory tracking of unmanned vehicle based on improved model predictive control

Wang Miao,Wu Yefu, Huang Fengyu
Institute of Computer Science and Technology,
Hubei Key Laboratory of Transportation Internet of Things Technology,
Wuhan University of Technology,
Wuhan, China
1819283069@qq.com,Wuyefu1988@qq.com,604767746@qq.com.

***Abstract*: Driverless technology has become a hot spot of current research, in order to improve the speed of unmanned vehicle path planning. Firstly, the artificial potential field (APF) model is added to the objective function of model predictive control (MPC) as a path planning term to make path planning and trajectory tracking simultaneously, so as to optimize the model controller.Secondly, by introducing the adjustment factor of objective function to solve the problem that MPC objective function may have no optimal solution, and then linearize the objective function to accelerate the speed of solving the objective function. Finally, Matlab / Simulink and CarSim are used for joint simulation experiments to compare the path planning and tracking effects before and after the improvement and optimization of the controller. The experimental results show that the improved model predictive controller has faster solution speed.**

Keywords: artificial potential field;model predictive control; unmanned vehicle; path planning; trajectory tracking;

I. INTRODUCTION

Most of the existing studies are based on APF for local path planning, and then MPC for trajectory tracking, but the real-time performance is not high[1].In view of this disadvantage, local path planning and trajectory tracking control are integrated to realize vehicle trajectory tracking at the same time.When designing the objective function of MPC, the APF environment model is added to the objective function as a path planning item, and the adjustment factor is added as an optimal solution guarantee item.Linearize the designed objective function to speed up its solution speed.Finally, through simulation experiments, the effects of path planning and trajectory tracking before and after controller improvement and optimization are compared.

II. TRAJECTORY TRACKING BASED ON MPC

Model predictive control (MPC) is an optimized control method, which has three main characteristics: predictive model, rolling optimization and feedback correction[2].

A. Vehicle dynamics model

Since the driving scene in this paper is a structured road and the vehicle speed is fast, vehicle dynamics is selected as the predictive model of MPC. The vehicle has three directions of motion, i.e. longitudinal, lateral and yaw motion[3]. Fig.1 is a vehicle monorail model used to describe vehicle dynamics.

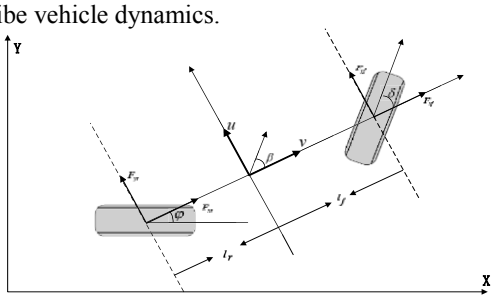


Figure 1. Vehicle monorail model.

According to the force condition of the vehicle, the vehicle dynamics model can be obtained by (1):

$$\begin{cases} ma_x = F_{xf} + F_{xr} \\ mv(\dot{\beta} + \dot{\varphi}) = F_{yf} + F_{yr} \\ I_z\ddot{\varphi} = l_f F_{yf} - l_r F_{yr} \\ \dot{X} = \dot{x}\sin\varphi - \dot{y}\sin\varphi \\ \dot{Y} = \dot{x}\sin\varphi - \dot{y}\sin\varphi \end{cases} \tag{1}$$

When the current wheel angle and tire longitudinal slip ratio are small, the tire force can be approximately described by linear function, so the longitudinal force and lateral force of tire can be obtained by (2):

$$\begin{cases} F_{xf} = C_{xf}s_f \\ F_{xr} = C_{xr}s_r \\ F_{yf} = C_{yf}\alpha_f \\ F_{yr} = C_{yr}\alpha_r \end{cases} \tag{2}$$

α_f , α_r ,represent the side deflection angle of front and rear wheels respectively, as shown in (3):

$$\begin{cases} \alpha_f = \delta - \dfrac{u}{v} - \dfrac{l_f\dot{\varphi}}{v} \\ \alpha_r = \dfrac{u - l_r\dot{\varphi}}{v} \end{cases} \tag{3}$$

F_{zf} , F_{zr} are the forces in the z-axis direction on the front and rear tires of the vehicle,then the forces on the tire in the Y direction can be expressed by (4):

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00021

$$\begin{cases} F_{yf}=C_{yf}\alpha_f=C_{yf}(\delta-\dfrac{u}{v}-\dfrac{l_f\dot{\varphi}}{v}) \\ F_{yr}=C_{yr}\alpha_r=C_{yr}(\dfrac{u-l_r\dot{\varphi}}{v}) \end{cases} \tag{4}$$

For the sum of forces in the z-axis direction on the front and rear tires of the vehicle, if the driving speed of the vehicle changes slowly, it can be calculated by (5):

$$\begin{cases} F_{zf}=\dfrac{l_r mg}{2(l_f+l_r)} \\ F_{zr}=\dfrac{l_f mg}{2(l_f+l_r)} \end{cases} \tag{5}$$

Based on the above discussion, the following nonlinear vehicle dynamics model can be obtained by (6):

$$\begin{cases} ma_x=C_{xf}s_f+C_{xr}s_r \\ mv(\dot{\beta}+\dot{\varphi})=C_{yf}(\delta-\dfrac{u}{v}-\dfrac{l_f\dot{\varphi}}{v})+C_{yr}(\dfrac{u-l_r\dot{\varphi}}{v}) \\ I_z\dot{\omega}=l_fC_{yf}(\delta-\dfrac{u}{v}-\dfrac{l_f\dot{\varphi}}{v})-l_rC_{yr}(\dfrac{u-l_r\dot{\varphi}}{v}) \\ \dot{X}=\dot{x}\sin\varphi-\dot{y}\sin\varphi \\ \dot{Y}=\dot{x}\sin\varphi-\dot{y}\sin\varphi \end{cases} \tag{6}$$

In the established vehicle dynamics model, $\chi=\left[a_x,\ \delta\right]^T$,is used to represent the selected longitudinal acceleration and the vehicle front wheel Angle position. (X,Y), φ, $\dot{\varphi}$, u ,and v respectively represent geodetic abscissa and ordinate, heading angle, heading angular velocity, lateral velocity and longitudinal velocity. Finally expressed as, $\xi=\left[X,Y,\varphi,\dot{\varphi},u,v\right]$.

B. Linearization of vehicle dynamics model

Due to the high requirement of real-time performance of the controller when the unmanned vehicle runs at high speed, linear time-varying model predictive control is adopted.

Firstly, the established vehicle dynamics model can be expressed by (7):

$$\dot{\xi}=g(\xi,\chi) \tag{7}$$

When Taylor expansion is carried out at any point, (ξ_r,u_r) ,and only the first-order term is retained, the following can be obtained by (8):

$$\dot{\xi}=g(\xi_r,\chi_r)+\left.\frac{\partial g}{\partial\xi}\right|_{\substack{\xi=\xi_r\\ \chi=\chi_r}}(\xi-\xi_r)+\left.\frac{\partial g}{\partial\chi}\right|_{\substack{\xi=\xi_r\\ \chi=\chi_r}}(\chi-\chi_r) \tag{8}$$

Equation (8) can be written as (9) by introduced Jacobian determinants $J_g(\xi)$ and $J_g(\chi)$ of function g:

$$\dot{\xi}=g(\xi_r,\chi_r)+J_g(\xi)(\xi-\xi_r)+J_g(\chi)(\chi-\chi_r) \tag{9}$$

At point, (ξ_r,χ_r) ,subtract equations (9) and (7) to obtain (10):

$$\dot{\xi}-\dot{\xi}_r=J_g(\xi)(\xi-\xi_r)+J_g(\chi)(\chi-\chi_r) \tag{10}$$

let $\dot{\tilde{\xi}}=\dot{\xi}-\dot{\xi}_r$, $A(t)=J_g(\xi)$, $B(t)=J_g(u)$, $\tilde{\xi}=\xi-\xi_r$, $\tilde{\chi}=\chi-\chi_r$ then equation (10) can be written as (11):

$$\dot{\tilde{\xi}}=A(t)\tilde{\xi}+B(t)\tilde{\chi} \tag{11}$$

Model predictive control is to predict the output in the next period of time based on the current time combined with the measured value, and repeat the above steps at the next time[4].The whole process is discrete step by step, and the prediction model (vehicle dynamics model) is continuous, so it is also necessary to discretize the vehicle dynamics model. Discretize equation (11) to obtain (12) and (13):

$$\tilde{\xi}(t+1)=A(t)\tilde{\xi}(t)+B(t)\tilde{\chi}(t) \tag{12}$$

$$\chi(t)=\chi(t-1)+\Delta\chi(t) \tag{13}$$

C. Objective function design

In order to consider factors such as vehicle stability and economic comfort, vehicle speed and control increment are added to the objective function, and in order to ensure that the objective function has an optimal solution, the optimal solution guarantee term is added. The objective function of single vehicle driving is designed as (14):

$$\begin{aligned} J(\xi(t),\chi(t-1),\Delta\psi(t))=&\sum_{i=1}^{N_p}\left\|U_{all}(t+i|t)\right\|_Q^2+\\ &\sum_{i=1}^{N_p}\left\|V(t+i|t)-V_{des}\right\|_R^2+\\ &\sum_{i=1}^{N_{c-1}}\left\|\Delta\chi(t+i-1|t)\right\|_S^2+\rho\varepsilon^2 \end{aligned} \tag{14}$$

The first part is the vehicle environmental potential field. By combining the discrete vehicle dynamics model and the control increment input sequence at time t, the potential field value of the vehicle at time t+i in the prediction time domain is calculated[5]. Through the potential field value, the vehicle risk is evaluated, and the point with small potential field value is selected to be added to the trajectory as the optimal trajectory point for prediction.

D. Objective function optimization

If the designed nonlinear objective function is used directly, the optimization speed of model predictive controller will be slow[6].Therefore, the objective function of the model predictive controller is linearized to improve its optimization speed.

At time $t=k$, Taylor expansion is performed as (15):

$$J(\xi(t),\chi(t-1),\Delta\psi(t))=J(\xi(k),\chi(k-1),\Delta\psi(k))+\left.\frac{\partial J}{\partial t}\right|_{t=k}(t-k) \quad (15)$$

let a $=J(\xi(k),\chi(k-1),\Delta\psi(k))$, and let B $=\left.\frac{\partial J}{\partial t}\right|_{t=k}$, $E(\xi(t),\chi(t-1),\Delta\psi(t))$ is introduced as the linearized objective function ,the Taylor expansion of the objective function can be written as (16):

$$E(\xi(t),\chi(t-1),\Delta\psi(t))=b\cdot t+(a-b\cdot k) \quad (16)$$

E. Controller constraint formulation

Since the objective function uses the control increment as the input, linear constraints are added to the control quantity itself and the control increment, that is, linear constraints are imposed on the longitudinal acceleration, a_x , longitudinal acceleration change rate, $\dot{a}_x$, front wheel steering angle , δ ,and front wheel steering angle change rate, $\dot{\delta}$ [7].

The values of the upper and lower bounds of the above related control quantities and increments can be shown in Tab.Ⅰ.

TABLE I. UPPER AND LOWER BOUND VALUES OF CONTROL QUANTITY AND INCREMENTAL CONSTRAINT PARAMETERS

parameter	minimum	Maximum
Longitudinal acceleration, a_x (M / S²)	-4	2
Longitudinal acceleration change rate, $\dot{a}_x$ (M / S³)	-25	25
Front wheel steering angle, δ (Rad)	-15	15
Front wheel angle change rate, $\dot{\delta}$ (Rad/S)	-9.4	9.4

In order to improve the stability of the vehicle and the universality of the controller, the constraints such as centroid deflection angle, vehicle adhesion coefficient, front and rear tire side deflection angle and vehicle speed shall also be considered. Assuming that the driving road is in good condition, the values of, β , κ , α_f , α_r , V are shown in Tab.Ⅱ.

TABLE II. VALUES OF UPPER AND LOWER BOUNDS OF, β , κ , α_f , α_r , V

parameter	minimum	Maximum
Centroid deflection angle, β (Rad)	-12	12
K	0.7	1
Sideslip angle of front and rear tires, α_f 、 α_r (Rad)	-2.5	2.5
Longitudinal speed, V (M / s)	0	30

Combined with the vehicle dynamics model, objective function and constraints, the trajectory planning and tracking of unmanned vehicle based on model predictive control can be described as the optimization problem shown in (17):

$$\begin{aligned}
&\min_{\Delta\psi(t)}\left\{E\left(\xi(t),\chi(t-1),\Delta\psi\right)\right\}\\
&s.t.\quad \chi_{\min}(t)\le\chi(t)\le\chi_{\max}(t)\\
&\Delta\chi_{\min}(t)\le\Delta\chi(t)\le\Delta\chi_{\max}(t)\\
&\xi(t+1)=A(t)\xi(t)+B(t)\chi(t)\\
&\chi(t)=\chi(t-1)+\Delta\chi(t)\\
&a_{x,\min}\le a_x\le a_{x,\max}\\
&\delta_{\min}\le\delta\le\delta_{\max}\\
&\dot{a}_{x,\min}\le\dot{a}_x\le\dot{a}_{x,\max}\\
&\dot{\delta}_{\min}\le\dot{\delta}\le\dot{\delta}_{\max}\\
&\beta_{\min}\le\beta\le\beta_{\max}\\
&\kappa_{\min}\le\kappa\le\kappa_{\max}\\
&\alpha_{\min}\le\alpha_f,\quad \alpha_r\le\alpha_{\max}\\
&0\le V\le 1.1\cdot V_{des}
\end{aligned} \quad (17)$$

After solving (17) in each control cycle, a series of control input increments, $\Delta\psi_t^*$,in the control time domain, N_c ,can be obtained by (18):

$$\Delta\psi_t^*=\left[\Delta u_t^*,\Delta u_{t+1}^*,\cdots,\Delta u_{t+N_c-1}^*\right]^T \quad (18)$$

When the first element of the control sequence is applied to the system as the actual control increment, it can be calculated by (19):

$$\Delta\chi(t)=\Delta\chi(t-1)+\Delta\chi_t^* \quad (19)$$

After entering the next control cycle, repeat the above process, so as to track the trajectory through cyclic iteration.

III. SIMULATION EXPERIMENT

By building CarSim, Matlab / Simulink joint simulation platform, it is mainly to verify that the solution speed of the model controller after optimizing the objective function is faster.

The structure diagram of the simulation experiment is shown in Fig.2. The S-function in the figure is the implementation function of local path planning and trajectory tracking.

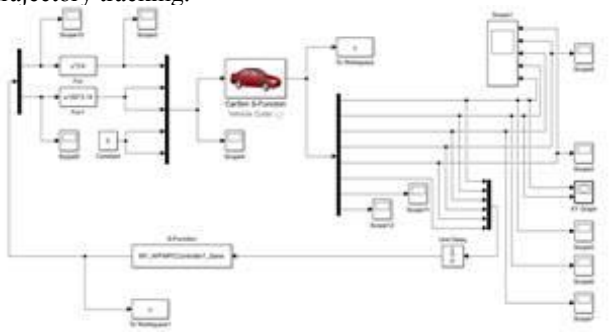


Figure 2. Structure of simulation experiment

In the experiment, eight simulation environments were designed according to the dynamic joining or leaving environment of environment vehicles, the existence of straddlable and unstraddlable obstacles on the lane, and whether the road was straight or curved.

Tab.III compares the time to solve the target solution before and after the improvement of MPC.

TABLE III. COMPARISON OF TIME CONSUMPTION BEFORE AND AFTER IMPROVING THE OBJECTIVE FUNCTION

Scenarios	**Time consuming(S)**	
	Before	***After***
Scenario 1	9.3	7.2
Scenario 2	7.8	6.2
Scenario 3	5.3	4.2
Scenario 4	5.3	4.1
Scenario 5	6.9	5.5
Scenario 6	4.2	3.4
Scenario 7	10.1	8.1
Scenario 8	9.9	7.9

It can be seen from the above table conclusion that by introducing APF and adjustment factor, the MPC controller improved by linearizing the objective function takes less time to solve the optimal path, which verifies that the solution speed of the optimized MPC is faster. At the same time, the experimental results also verify that the improved controller is more accurate in path planning and tracking, and the planned path is better.

IV. CONCLUSION

The optimized model controller of APF and MPC designed in this paper verifies that the unmanned vehicle can ensure the speed and tracking accuracy in the scene with interference vehicles and obstacles, and can realize the following, obstacle avoidance, lane changing and overtaking according to the road conditions.The disadvantage is that the driving environment of unmanned vehicles in this paper is a structured road, but for complex road scenes, the influencing factors in the driving environment need to be further considered.

REFERENCES

[1] ZHOU L H, SHI P L, JIANG J X. Simulation research on vehicle stability control based on collision avoidance trajectory tracking [J]. Journal of Shandong University of Technology (NATURAL SCIENCE EDITION), 2021,35 (05): 75-81

[2] LU H G, ZHAO S E. Research on intelligent vehicle trajectory tracking control based on robust Model Predictive Control [J / OL]. Journal of system simulation: 1-10 [2021-06-26].

[3] JI C K, DONG S P, MAO T L. Obstacle avoidance path planning algorithm based on Model Predictive Control[C]. 18th International Conference on Control, Automation and Systems (ICCAS), 2018: 141-143.

[4] JIE C L, Paramsothy J, Jeffrey L S. Combined speed and steering control in high-speed autonomous ground vehicles for obstacle avoidance using Model Predictive Control[J]. IEEE Transactions on Vehicular Technology, 2017, 66(10): 8746-8763.

[5] ZOU Kai, CAI Y F, CHEN L. Trajectory tracking method of unmanned vehicle based on incremental linear Model Predictive Control [J]. Automotive technology, 2019, (10): 1-7.

[6] WANG D, LI C H, GUO N, GAO T T, LIU G M. Local path planning of mobile robot based on improved Artificial Potential Field method [J]. Journal of Shandong University of Technology (NATURAL SCIENCE EDITION), 2021,35 (03): 1-6.

[7] Yadollah R, Amir K, Shih K C. A potential field-based model predictive path-planning controller for autonomous road vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(5): 1255-1267.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

NDT-based robot positioning system for large scale diversity environment

Qi Min[1] , Niansheng Chen[1]*, Guangyu Fan[1], Lei Rao[1]
[1]Shanghai DianJi University
Shanghai, China
hellominqi@163.com; chenns@sdju.edu.cn

Zhaohui Xu[2], Yiping Ma[2]
[2]AVIC Huadong Photoelectric Shanghai Co., Ltd.
Shanghai, China
yiping_ma@foxmail.com

***Abstract*—High accuracy positioning is an important challenge for robots in large scale environment. In order to reduce the localization errors in the environments, an efficient Normal Distributions Transform (NDT) localization method with multi-sensor fusion data fusion, namely FPCR-NDT-localization, is proposed. Firstly, the laser point cloud is pre-processed to remove the ground point cloud and the laser feature point cloud is extracted to reduce the laser point cloud alignment. Secondly, the inertial measurement unit pre-integration results are used to estimate the LIDAR inter-frame state transfer volume, and the point cloud alignment efficiency is accelerated by the feature point-based FPCR-NDT alignment method. Finally, the IMU pre-integration results and the LIDAR inter-frame alignment are fused to estimate the robot pose in the global map. In the experiments, the localization performance of HDL-localization, NDT-localization and FPCR-NDT-localization systems are tested using MulRan dataset. The results show that FPCR-NDT-localization has higher localization accuracy in different scenes and higher real-time performance compared to the original algorithm.**

Keyword—Mobile Robot; Normal Distribution Transformation; Positioning Technology; Point cloud Registration

I. INTRODUCTION

Autonomous navigation is one of the basic tasks of mobile robots[1]. In order to meet the challenges of large outdoor scene environment range and complex environmental features, it is necessary to improve the positioning accuracy and real-time performance in the process of navigation in large outdoor scenes. The positioning and navigation solutions in large scenarios mainly face the following problems. Firstly, the environment range is large and the accuracy of inter-frame pose estimation is insufficient, which leads to the accumulation of errors during the movement. Secondly, the real-time pose estimation during high-speed movement is insufficient, which leads to the decrease of positioning accuracy. Thirdly, the deflection angle of the robot is large at adjacent moments during the movement, which makes the existing inter-frame laser point cloud matching technology unable to be aligned.

II. RELATED WORK

Various solutions have been proposed to achieve localization targets and improve localization accuracy of mobile robots in large scenarios[2-5].The paper[4] proposes a positioning system based on the fusion of GNSS, LIDAR(Light Detection and Ranging), and IMU(Inertial Measurement Unit), which uses LIDAR intensity and height information to improve the accuracy and robustness of the positioning system. In this system data fusion framework Import real-time laser odometer, and use the error state Kalman filter to fuse the positioning results of different sensors. Later, the paper [5] improved the data fusion framework of [4] by introducing a factor graph optimization based approach and also adding a scene change detection component. This method makes better use of multi-sensor data, however, it relies too much on the smoothness of the sensor acquisition data.

LIDAR-based localization technology mainly realizes localization through inter-frame matching of laser point cloud data and local map matching[6-7]. During robotic LIDAR movement, point cloud matching techniques are used to determine the relative transformation relationship between two frames of the point cloud, which helps in robot odometry estimation in unknown environments[8]. Moreover, precise transformation parameters can determine the change in attitude between two moments [9]. The NDT algorithm based on the local alignment method treats the point cloud as a set of Gaussian distributions, which is applied to a statistical model of 3D points, and determines the matching parameters between the point clouds using standard optimization methods[10]. Peter et al. first used the point cloud of the current scan frame to construct a normal distribution of 2D variables for 2D data alignment[11]. Magnusson et al. applied this method to the scanning alignment process of 3D LIDAR data [12]. In large scale outdoor environment, the number of point clouds is huge and the environment is variable, so the efficiency and accuracy of point cloud alignment are reduced, to solve such problems, the feature-based laser point cloud alignment method is proposed[13]. This method of quickly aligning two frames of point clouds by features has efficient feature extraction performance and is suitable for large-scale outdoor scene feature extraction.

For application areas such as outdoor laser point cloud alignment and robot localization in large scale environment, the research of NDT matching technology based on effective point cloud feature extraction method is still meaningful. Based on the shortcomings of existing positioning systems applied to outdoor large-scale environments, the following work is made in this paper. (1) Reducing the amount of laser point cloud alignment and improving the computational efficiency by ground point filtering of the original laser point cloud.(2) Propose a Feature point cloud registration NDT algorithm, FPCR-NDT, for large-area scenes based on the original NDT method and the study of feature extraction-based LIDAR point cloud alignment.(3) Propose a lightweight multi-sensor data fusion localization system that fuses IMU and laser odometer sensor information and incorporates a local point cloud map alignment step to achieve high accuracy localization.

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00022

III. POSITIONING SYSTEM FRAMEWORK

The corresponding schematic diagram of the positioning system is shown in Fig. 1. Firstly, in the laser point cloud pre-processing stage, non-ground point clouds are segmented to improve subsequent alignment efficiency by filtering ground point cloud information for concise processing; then, the point cloud is roughness classified and the planar feature points are extracted from it for subsequent alignment; meanwhile, the IMU odometry information is pre-integrated to predict the position transformation within two frames; in the measurement update stage, the IMU predicted position is used as the iterative initial value for the laser odometry. In the measurement update stage, the IMU predicted pose is used as the iterative initial value for the inter-frame NDT iterative alignment of the laser odometry; finally, the point cloud pose derived from the laser odometry is aligned with the local point cloud map to update the robot's own pose and achieve positioning.

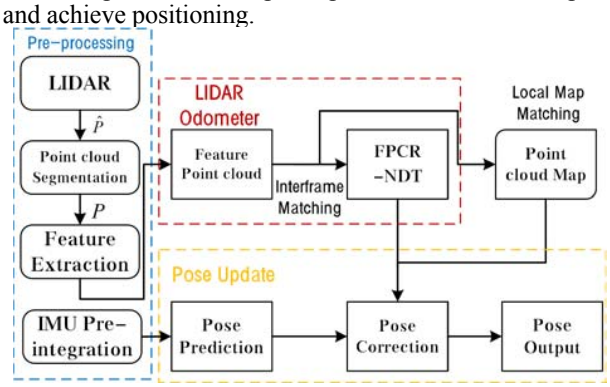


Figure 1. Schematic diagram of IMU fusion localization system based on FPCR-NDT matching algorithm

IV. 3D POINT CLOUD MAP POSITIONING SYSTEM

A. Ground Point cloud Segmentation

In the point cloud obtained by LIDAR scanning, because the ground point cloud data contains few effective features and will have certain influence on the subsequent classification and recognition of obstacles, and a large amount of redundant data on the ground will reduce the operation efficiency of the matching algorithm, so it is necessary to filter the ground part of the point cloud in the pre-processing stage of the point cloud information. In order to filter out the ground points in the point cloud, a fast ground segmentation method is used to segment the ground point cloud information. The filtered ground point cloud can effectively reduce data redundancy, reduce memory consumption.

B. Problem Description

During the robot movement, the laser odometer converts the position of the two point clouds by matching the point clouds, and the position transformation matrix of the robot during the acquisition time of the two point clouds is derived. Define the point cloud as the set $P=\{p_i \in R^3 | i=1,2\ldots,n\}$, where p_i consists of three coordinate components, $p_i=\{p_i^x, p_i^y, p_i^z\}$. In the three-dimensional space, six-dimensional vectors are defined to encode the position transformation parameter ε. Also, the robot state variable X is defined.

$$\begin{cases} \varepsilon = \left[t_x, t_y, t_z, \phi, \theta, \psi \right] \\ X = \left[x, y, z, \alpha, \beta, \gamma \right] \end{cases} \tag{1}$$

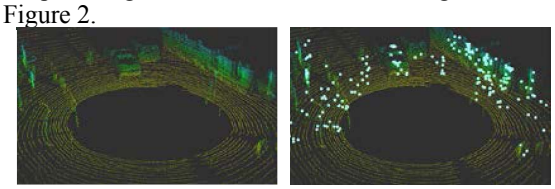
Using the Euclidean sequence z-y-x, the three-dimensional transformation function *Trans*( ) applying the transformation parameter ε is defined in (2), where the transformation parameter ε consists of the rotation parameter R and the translation parameter T.

$$Trans(\varepsilon, p) = \left[R_\phi \quad R_\theta \quad R_\psi \right] p + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \tag{2}$$

Where R_ϕ is the rotation matrix when rotating around the x-axis at angle ϕ and $T=[t_x, t_y, t_z]^T$ is the translation vector between the origin of the two coordinate systems. In this paper, the current point cloud data to be aligned is defined as the scan frame and the completed aligned point cloud data is defined as the reference frame.

C. Feature Point Extraction

In order to improve the alignment accuracy and real-time performance, the initial point cloud is classified and features are extracted to improve the efficiency of subsequent alignment, and the schematic diagram is shown in Figure 2.



(a) Original Point Cloud (b) Feature point extraction

Figure 2. Extraction of feature points from the original point cloud

The feature points are extracted for the point cloud in Fig. (2)a using a planar feature detection based approach, so that M is taken as the point set of consecutive points x_i in the same line, where x_i is at the center of M, as shown in Fig. 3.

The roughness r of each point is calculated in (3), where $|M|$ is the number of point clouds in the set, and in this paper $|M|=15$. $X_{(k,j)}$ is the coordinates of the adjacent points to the left and right of point x_i.

$$r = \frac{1}{|M| \cdot \| X_{(k,i)}^L \|} \| \sum_{j \in M, j \neq i} (X_{(k,i)}^L - X_{(k,j)}^L) \| \tag{3}$$

Set the roughness threshold as R, traverse all x_i to consider the points with roughness less than R as feature points, as shown in Fig. (2)b.

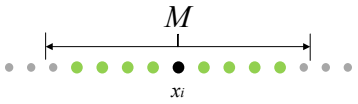


Figure 3. Single line continuous point cloud distribution

D. FPCR-NDT Point Cloud Matching Algorithm

Point cloud alignment is performed for the scanned frames and reference frames for which feature point extraction has been completed, as follows.

Step1. For the current scanned frame feature point cloud set P^S and reference frame point cloud set P^R, the feature point clouds of scanned frame and reference frame are divided into two sets of raster C^S, C^R, $P^S=\{C_1^S, C_2^S,\ldots, C_n^S\}$, $P^R=\{C_1^R, C_2^R,\ldots, C_n^R\}$,where n is the number of feature point cloud raster cells, and the feature points contained in

the raster cells generate a Gaussian distribution $N(u_i^s, \Sigma_i^s)$,$N(u_i^R, \Sigma_i^R)$.

Step2. Initialize the positional transformation parameter $\varepsilon=[t_x, t_y, t_z, \phi, \theta, \psi]$.

Step3. The alignment of the raster distribution to the inter-distribution is performed, and the inter-distribution alignment performance is judged by the following calculation results.

$$L_2 = \sqrt{\sum_{i=1}^{n}(X_i^S - X_i^R)^2} \tag{4}$$

$$\gamma_{ij} = \left[Trans(\varepsilon, u_i^S) - u_j^R \right] \tag{5}$$

The next step defines the alignment error function Ψ( ), which is used to calculate the alignment error of this parameter applied to the two-frame point cloud, as in (6).

$$\Psi(\varepsilon) = -\sum_{i=1}^{|C^S|}\sum_{j=1}^{|C^R|} \exp(-\frac{1}{2}\gamma_{ij}^T[\varepsilon_s^T \Sigma_i^R \varepsilon_s + \Sigma_j^S]^{-1}\gamma_{ij}) \tag{6}$$

Step4. The alignment error function obtained in the previous step is optimized and its gradient vector is solved by components. Let α and β be one of the components of the six-dimensional vector ε, respectively[14]. The partial derivative of this component α in ε is solved as in (7).

$$\frac{\partial}{\partial\alpha}\Psi(\varepsilon) = \frac{1}{2}(\gamma_{ij}^T B \boldsymbol{j_a} - \gamma_{ij}^T B Z_a B \gamma_{ij}) \exp(-\frac{\gamma_{ij}^T B \gamma_{ij}}{2}) \tag{7}$$

where B, $\boldsymbol{j}_\alpha$, and Z_α are the calculated process variables as shown in (8), (9), and (10).

$$B = (R^T \Sigma_i R + \Sigma_j)^{-1} \tag{8}$$

$$\boldsymbol{j_a} = \frac{\partial}{\partial\alpha}(Ru_i + t - u_j) \tag{9}$$

$$Z_a = \frac{\partial}{\partial\alpha}(R^T \Sigma_i R) \tag{10}$$

Solve the alignment error function for component β and calculate the Hessian matrix.

$$\frac{\partial^2}{\partial\alpha\partial\beta}\Psi(\varepsilon) = d_1 d_2 (\boldsymbol{j_a}^T B \boldsymbol{j_a} - 2\gamma_{ij}^T B Z_a \boldsymbol{j_a} + \gamma_{ij}^T B H_{ab} - \gamma_{ij}^T B Z_a B Z_b B \gamma_{ij} - \frac{1}{2}\gamma_{ij} B Z_{ab} B \gamma_{ij} - \frac{d_2}{4}\boldsymbol{q}^T \boldsymbol{q}) \exp(-2\frac{d_2 \gamma_{ij}^T B \gamma_{ij}}{2}) \tag{11}$$

where H_{ab}, Z_{ab}, and $\boldsymbol{q}$ are process variables as shown in (12), (13), and (14).

$$H_{ab} = \frac{\partial^2}{\partial\alpha\partial\beta}(Trans(\varepsilon, u_i^S) - u_j^R) \tag{12}$$

$$Z_{ab} = \frac{\partial^2}{\partial\alpha\partial\beta}(R^T \Sigma_i R) \tag{13}$$

$$\boldsymbol{q} = \gamma_{ij}^T B \boldsymbol{j_a} - \gamma_{ij}^T B Z_a B \gamma_{ij} \tag{14}$$

For each component α, β of the six-dimensional transformation parameter ε, the first-order derivative and second-order derivative of the error function about the components are solved, and the partial derivative expression about the variables is derived, and the alignment error function is minimized and solved, as shown in (15) and (16) to obtain the final transformation parameter ε^* of the two-frame point cloud, at which point the robot carrier state variable X_{scan} be derived as follows.

$$\varepsilon^* = \arg\min \Psi(\varepsilon) \tag{15}$$

$$X_{scan} = Trans(\varepsilon^*, X_{ref}) \tag{16}$$

E. LIDAR/IMU Combined Positioning System

The combined positioning method consists of the following steps: first, the IMU integration information is used to derive the inter-frame pose offset of the robot for coarse positioning of the robot, and this pose is imported into the FPCR-NDT matching as the initial value for the iterative alignment of the reference frame and the scan frame. After the two frames are aligned, the observed scanned frames are finely aligned with the local point cloud map, and the robot's pose in the map is obtained by using its state update. The state transfer process is defined as a discrete-time process model, and the FPCR-NDT point cloud alignment and local map matching are incorporated into the state correction process.

In the attitude prediction stage, the robot attitude information is derived from the IMU to provide initial values of the attitude, LIDAR point cloud alignment and other observation data. Set robot state $X_t = [x_t, y_t, z_t, \alpha_t, \beta_t, \gamma_t]$ at moment t during the time period in which the two state nodes (L_t, L_{t+1}) are located , the local map alignment step as a positional correction by the relative positional inferred transformation matrix T_{IMU} of IMU with the alignment relationship $T_{point2point}$ of the point cloud of the two preceding and following frames. Let the alignment relationship between the current frame and the local map be $T_{point2map}$, and derive the robot pose prediction value X'_{t+1} with the updated value X_{t+1} at moment $t+1$.

$$X'_{t+1} = T_{point2point}(T_{IMU}(X_t)) \tag{17}$$

$$X_{t+1} = T_{point2map}(X'_{t+1}) \tag{18}$$

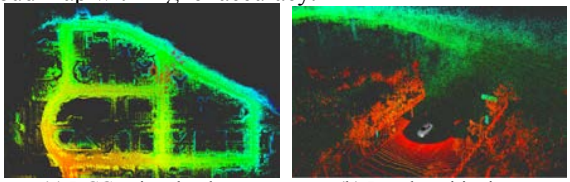
V. EXPERIMENT AND ANALYSIS

In order to validate the localization performance of the proposed localization method, experiments are conducted to compare with existing open source solutions HDL-localization[15] with NDT-localization. In this paper, we use the mulran dataset to evaluate the localization accuracy, trajectory and other metrics of our method.

A. Experimental Settings

First, SC-LeGO-LOAM[16], an open source 3D laser slam solution with excellent performance at present, is used to generate a high-precision point cloud map. Then, the point cloud map is imported for initialization, and the robot conducts navigation experiments in this point cloud map.

Fig. 4 shows the DCC scene in the dataset. From the figure, it can be obtained that the current frame range of the point cloud is larger and can match with the local point cloud map with higher accuracy.



(a) DCC point cloud map (b) Local positioning

Figure 4. DCC Scene (a) shows the point cloud map of the site, and Fig. (b) shows the effect of the robot driving in the point cloud map

B. Trajectory Error Analysis

To test the performance of the FPCR-NDT positioning system, the trajectories of the above positioning system traveling in the dataset DCC scenario were derived using the trajectory evaluation tool evo, as shown in Figure 5.

From the partially enlarged area of Fig. 5, it can be concluded that among the three sets of trajectories, the trajectories of the FPCR-NDT positioning system are closer to the real motion trajectories, and most of the positioning achieves decimeter-level accuracy compared with the real trajectories, and there is no obvious cumulative error during the driving process.

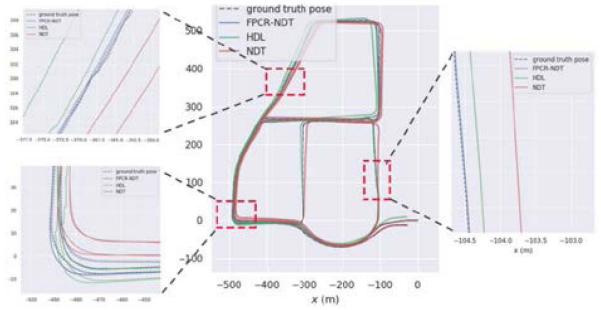
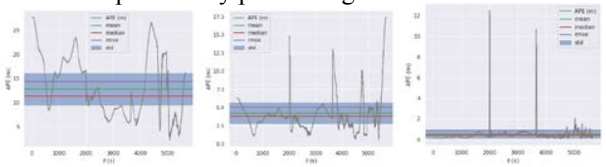


Figure 5. Comparison of three positioning methods in DCC scene motion trajectory

Figure 6 shows the comparison of the absolute positional error between the motion trajectory and the real trajectory obtained by the above three positioning systems in the DCC scenario. Table 1 shows the absolute error information for the DCC, KAIST and Riverside scenes in the Mulran dataset. In the moving process, HDL-localization and NDT-localization localization systems, which only rely on point cloud alignment for pose estimation, can hardly guarantee that the error is within a small range, and the FPCR-NDT localization system provides better pose estimation for the robot through IMU at the front end, which provides iterative initial values for subsequent alignment while ensuring computational efficiency and avoiding The iterative process falls into the local optimal situation, which is reflected in the higher accuracy and smaller root-mean-square error and standard deviation of the trajectory of the LIDAR and IMU data fusion complementary positioning method.



(a) NDT (b) HDL (c) FPCR-NDT

Figure 6. Absolute error of trajectory positioning of DCC scenes

VI. CONCLUSION

In order to improve the positioning accuracy for moving robots in large scale and diverse environments, a real-time localization system based on multi-sensor combination localization with feature point cloud alignment is proposed. In the system, the laser point cloud is pre-processed to remove the ground point cloud and the laser feature point cloud is extracted to reduce the laser point cloud alignment. And the inertial measurement unit pre-integration results are used to estimate the LIDAR inter-frame state transfer volume. Moreover, the LIDAR inter-frame alignment and the IMU pre-integration results are fused to estimate the robot pose in the global map. The experiments results show that FPCR-NDT-localization has higher accuracy and computing efficiency, compared to the original positioning system.

TABLE I. ABSOLUTE ERROR COMPARISON

Scene	method	mean	median	rmse	std
DCC	NDT	12.7691	11.4212	14.343	6.5164
	HDL	4.262	3.8684	5.092	2.7871
	FPCR-NDT	**0.565**	**0.4475**	**0.949**	**0.7763**
KAIST	NDT	15.195	13.6896	11.337	8.7607
	HDL	6.214	5.5518	4.366	3.7627
	FPCR-NDT	**0.832**	**0.7710**	**1.189**	**1.5937**
Riverside	NDT	16.972	16.7324	15.335	13.109
	HDL	7.185	8.6252	6.377	5.8187
	FPCR-NDT	**1.267**	**0.9018**	**1.792**	**1.8676**

REFERENCES

[1] Lambrecht, J. , and E. Funk . Edge-Enabled Autonomous Navigation and Computer Vision as a Service: A Study on Mobile Robot's Onboard Energy Consumption and Computing Requirements. 2020.

[2] Bachrach, A. G. , et al. "RANGE - robust autonomous navigation in GPS-denied environments." 2010 IEEE International Conference on Robotics and Automation IEEE, 2010.

[3] Anderson, D. M. , R. E. Estell , and T. S. Schrader . "Calculating Foraging Area Using Global Navigation Satellite System (GNSS) Technology." Rangelands 36.6(2014):31-35.

[4] Wan, G. , et al. "Robust and Precise Vehicle Localization based on Multi-sensor Fusion in Diverse City Scenes." (2017).

[5] Ding, W. , et al. "LIDAR Inertial Odometry Aided Robust LIDAR Localization System in Changing City Scenes." 2020 IEEE International Conference on Robotics and Automation (ICRA) IEEE, 2020.

[6] Wolcott, R. W. , and R. M. Eustice . "Visual localization within LIDAR maps for automated urban driving." IEEE(2014).

[7] Durrantwhyte, H. F. , and T. Bailey . "Simultaneous localization and mapping (SLAM): part II." IEEE Robotics & Amp Amp Automation Magazine 13.2(2006):99 - 110.

[8] Gressin, A. , et al. "Towards 3D LIDAR point cloud registration improvement using optimal neighborhood knowledge." ISPRS Journal of Photogrammetry and Remote Sensing 79.may(2013):240-251.

[9] Das, A. , J. Servos , and S. L. Waslander . "3D scan registration using the Normal Distributions Transform with ground segmentation and point cloud clustering." 2013 IEEE International Conference on Robotics and Automation IEEE, 2013.

[10] M Magnusson, M Magnusson, and M Magnusson. "Title: The Three-Dimensional Normal-Distributions Transform — an Efficient Representation for Registration, Surface Analysis, and." (2014).

[11] Biber, Peter . "The normal distributions transform: a new approach to laser scan matching." 2003.

[12] Magnusson, M. , A. Lilienthal , and T. Duckett . "Scan registration for autonomous mining vehicles using 3D‐NDT." Journal of Field Robotics 24.10(2007):803-827.

[13] Ji, Z. , and S. Singh . "Visual-LIDAR odometry and mapping: low-drift, robust, and fast." IEEE International Conference on Robotics & Automation IEEE, 2015.

[14] "Fast and Accurate Scan Registration through Minimization of the Distance between Compact 3D NDT Representations." International Journal of Robotics Research 31.12(2012):1377-1393.

[15] "A Portable 3D LIDAR-based System for Long-term and Wide-area People Behavior Measurement." International Journal of Advanced Robotic Systems 16.2(2019).

[16] Kim, G. , and A. Kim . "Scan Context: Egocentric Spatial Descriptor for Place Recognition Within 3D Point Cloud Map." 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) IEEE, 2018.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Research on Detection Method of Abnormal Trajectory of Port Operation Vehicles

Shuaihui Li
School of Computer Science and Technology,
Wuhan University of Technology
Wuhan, China
Email: wiki_hui@qq.com

Hongxing Liu,Hanbing Yao
Hubei Provincial Key Laboratory of Transportation
Internet of Things Technology
Wuhan, China
Email: liuhongxing@whut.edu.cn,
yaohb@whut.edu.cn

***Abstract*—Port operation vehicles are mainly responsible for the transshipment of goods. If there is inadequate supervision in the process of transshipment it is easy to cause such behaviors as cargo leakage, cargo theft and illegal parking of drivers, which cause economic losses to the port. In order to discover such behaviors in time, the unsupervised anomaly detection algorithm Self-encoder-based Deep Feature Fusion Model(S-DFFM) is proposed to judge whether the trajectory of operation vehicles is abnormal or not. The method comprehensively characterizes the trajectory by fusing the shallow features of low-dimensional trajectory and the deep features of high-dimensional trajectory, which frees the trajectory from the limitation of spatial attributes. The experimental data adopts the real trajectory data of one month (7,547 trips) of operating vehicle trajectories of a port in Chongqing, and the experimental results show that S-DFFM can better represent the trajectory features, and the accuracy of trajectory abnormality detection using S-DFFM is as high as 96%.**

***Keywords*—Trajectory; Anomaly detection; Feature fusion; Unsupervised detection;**

I. Introduction

The port is the hub center of water and land transportation, and the loading and unloading operation of the port is usually the transit of goods by operating vehicles. In the process of transshipment, due to the complex roads in the port, various operation types, harsh yard environment, and uncontrollable human factors, the process is prone to some abnormal operation behaviors such as cargo leakage (cargo not passing through the weighing equipment), cargo theft (transferring cargo from owner A to owner B's yard), and illegal parking (eating snack during operation). The operating vehicle will generate massive trajectory data in the process of operation, and by analyzing the trajectory data[1] and detecting the abnormal trajectory, the real behavior pattern of the operating vehicle hidden in the normal operation mode can be found, and the abnormal intention of the operating vehicle operation can be exposed, so that the supervisors can make corresponding measures.

Knorr et al[2] transformed the trajectory into some representative features consisting of position, direction and velocity, instead of just viewing the trajectory as a series of points, and then later detected the trajectory anomaly by comparison of distances. This method only compares the overall characteristics of the trajectory and ignores the local characteristics, and the problems exposed by this method become more and more obvious as the trajectory becomes longer. To solve the problems caused by comparing whole trajectories, Lee et al[3] designed a division and detection framework and proposed the TRAOD(Trajectory Outlier Detection) algorithm accordingly.Zhu et al[4] proposed a time-dependent popular path trajectory outlier detection algorithm TPRO(Time-dependent Popular Routes), which takes into account temporal anomalies but requires a lot of time computation during preprocessing.Laxhammar et al[5] proposed an online supervised trajectory anomaly detection algorithm SHNN-CAD(Sequential Hausdorff Nearest-Neighbor Conformal Anomaly Detector), which applies a consistent variance detector[6] to calculate the statistical confidence values of trajectories.

In this paper, considering the limitations of existing distance or density-based algorithms and the special characteristics of in-port trucking operations (some road networks are dense, operation types and operation goods are diverse, there are stationary and mobile states in operation links, and operation drivers can be changed, etc.),we propose the Self-encoder-based Deep Feature Fusion Model (S-DFFM), which converts shallow feature sequences into deep feature sequences and splices shallow feature sequences with deep feature sequences into fused feature sequences to comprehensively represent the features of trajectory segments from both deep and shallow layers.In the anomaly detection stage, the anomalous trajectory segments are extracted by comparing the similarity indices between the fused feature sequences, and then the percentage of the anomalous trajectory segments on the trajectory as a whole is considered to finally detect the anomalous trajectory.

II. Operating Vehicle Trajectory Feature Extraction And Conversion

Deep neural networks have been shown to have a strong ability to learn verbal data[7], textual data[8], and image data[9]. Trajectory data is three-dimensional data consisting of two-dimensional coordinates and timestamps, and it has been shown through empirical studies that it is not feasible to simply input three-dimensional data of the original trajectory into

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00023

deep learning algorithms[10]. Therefore, in order to make the trajectory data suitable for deep neural network algorithms, trajectory features with stronger semantic expressions need to be extracted from the original trajectory data and the original 3D trajectory data is replaced by a representation of feature sequences.

As shown in Fig.1, the core of feature extraction and transformation is divided into two steps. First, the geometric features are extracted from a number of original trajectory fragments of variable length after division, and the geometric features are transformed into a shallow feature sequence F of fixed length after dimensional expansion; second, the shallow feature sequence is transformed into a fused feature sequence FM using a deep feature fusion model based on self-encoder, i.e., the shallow feature sequence F is stitched with the deep trajectory feature sequence FD to obtain the fused feature sequence FM.

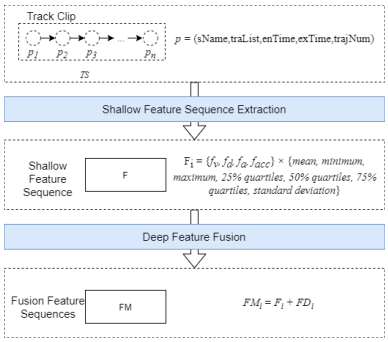


Fig. 1. Feature Extraction and Transformation.

A. Shallow Feature Extraction

Definition 1-1. Shallow feature sequence: the geometric features of the expanded trajectory are called the shallow feature sequence of the trajectory, and the set Q composed of n shallow trajectory feature sequences is called the set of shallow feature sequences, denoted as $Q = \{F_1, F_2, F_3 \dots F_n\}$,where each shallow feature sequence F_i represents a trajectory segment $TS_i (1 \leq i \leq n)$.The dimensionality expansion requires the calculation of mean, minimum, maximum, 25%, 50% and 75% quartiles, and standard deviation[11] for each trajectory segment feature, and the obtained calculation result is the shallow feature sequence of each trajectory segment.According to Definition 1-1, the velocity change f_v, distance change f_d, angle change f_a and acceleration change f_{acc} of the trajectory analyzed in this section, the shallow feature sequence F is expressed as $F = \{f_v, f_d, f_a, f_{acc}\} \times \{$ mean, minimum, maximum, 25% quartiles, 50% quartiles, 75% quartiles, standard deviation $\}$, It is obvious that the shallow feature sequence F is a $28(4 \times 7)$ dimensional feature sequence. The shallow feature sequence can visually reflect the trajectory fragments and quantify the original GPS trajectory data at the same time. The set of shallow feature sequences Q is used to represent the trajectory fragment set TS, which is convenient for subsequent analysis and processing.

B. Deep Feature Fusion

After obtaining the shallow feature sequence, the depth feature fusion model based on the self-encoder is used to convert the shallow feature sequence into the deep feature sequence, and the shallow feature sequence is stitched with the deep feature sequence into the fused feature sequence to comprehensively represent the features of the trajectory segments from both deep and shallow aspects.

1) Self-Encoder: Fig.2 shows a three-layer self-encoder model, where L_1 is the input layer, L_2 is the hidden layer, and L_3 is the output layer. The input layer L_1 and the hidden layer L_2 constitute the encoder, which is responsible for extracting potential features. The hidden layer L_2 and the output layer L_3 form the decoder, which is responsible for reconstructing the data input from the hidden layer. Denote the data in input layer L_1, hidden layer L_2 and output layer L_3 by x, h and y $(x, y \in [0,1]^n, \quad h \in [0,1]^m)$. W is the weight matrix$(W \in R^{m \times n})$, b_1 $(b_1 \in R^m)$ denotes the bias of the input layer L_1 and b_2 $(b_2 \in R^n)$ denotes the bias of the hidden layer L_2.

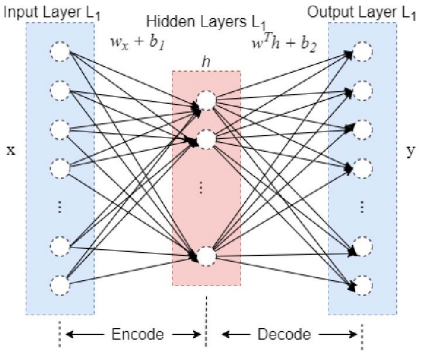


Fig. 2. Self-Encoder.

The value h in the hidden layer L_2 and the value y in the output layer L_3 are calculated by equation (1) and equation (2), where $\sigma()$ is the sigmod activation function.

$$h = f(x) = \sigma(Wx + b_1) \tag{1}$$

$$y = g(h) = \sigma\left(W^T h + b_2\right) \tag{2}$$

The purpose of using the self-encoder is to extract the salient high semantic features of the original data from the hidden layer, while having the function of dimensionality and data noise reduction. In order to make the learned hidden layer meaningful, the data output from the encoder needs to enter the decoder again for decoding, and through repeated iterations, the parameters of input and output are updated so that the values of y and x are close to each other, i.e., the self-encoder tries to learn the target optimization function of $y = F_{w.b}(x) \approx x$. The target optimization function can be expressed by the reconstruction error function $L(x, y)$, as

shown in equation (3). According to the reconstruction error function, the loss functionloss can be obtained, and the $loss$ function is calculated by equation (4).

$$L(x, y) = -\sum_{i=1}^{n} (x_i \log(y_i) + (1 - x_i) \log(1 - y_i)) \quad (3)$$

$$\text{loss} = \sum_{x \in Q} L(x, g(f(x))) \quad (4)$$

The weight matrix W and the biases b_1 and b_2 are updated as shown in equation (5), equation (6) and equation (7). Where,α denotes the learning rate.

$$W^{\text{new}} = W^{\text{old}} - \alpha \frac{\partial loss}{\partial W^{\text{old}}} \quad (5)$$

$$b_1^{new} = b_1^{\text{old}} - \alpha \frac{\partial loss}{\partial b_1^{\text{old}}} \quad (6)$$

$$b_2^{\text{new}} = b_2^{\text{old}} - \alpha \frac{\partial \text{loss}}{\partial b_2^{\text{old}}} \quad (7)$$

2) Fusion Feature Sequence Conversion: Using the deep feature fusion model based on self-encoder, the shallow feature sequences and deep feature sequences are transformed into fused feature sequences, which takes into account the detailed information of the shallow features and combines the semantic information of the deep features, and can highlight both the shallow and deep features of the trajectory segments, and the model structure is shown in Fig.3.

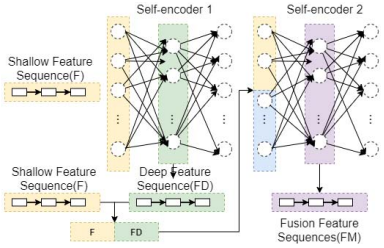


Fig. 3. Deep Feature Fusion Model.

The dimension of the shallow feature sequence at the input is 28. Considering the calculation time consumption and the size of the reconstruction error, the dimensions of the hidden layer L_2 of the self-encoder 1 and the self-encoder 2 are set to 10 and 15 respectively. When the 28-dimensional F is input to the model, the 10-dimensional FD is obtained through auto-encoder 1, and then spliced with F and sent to auto-encoder 2, and finally 15-dimensional FM is obtained.

C. Unsupervised Learning Based Feature Sequence Clustering

DBSCAN(Density-Based Spatial Clustering of Applications with Noise) can find clusters of arbitrary shapes and irregular clusters without prior knowledge of the number of clusters, but DBSCAN is too sensitive to parameters, and the clustering effect may be completely different under different parameters, which poses a great challenge to the parameter adjustment in experiments.OPTIC(Ordering Points to Identify the Clustering Structure) is an improved algorithm of DBSCAN, which can effectively solve the parameter sensitivity problem.

1) Similarity Index Anomaly Detection: Definition 1-2. Similarity index: The similarity matrix S is obtained by fusing the cosine similarity between each element of the set of feature sequences, and the set consisting of the similarity indices of all trajectory segments is the similarity indices A(The sum of the i-th row of S $ano_i = \sum s_i (1 \leq i \leq n)$)

Definition 1-3. anomalous trajectory fragment: given a similarity index threshold ρ, for a trajectory fragment TS, if its similarity index is less than the threshold ρ, then the trajectory is said to be fragment TS is an anomalous trajectory fragment.

Definition 1-4. Abnormal trajectory: Given an abnormal length ratio threshold τ, a trajectory T is said to be abnormal if the length ratio of its abnormal trajectory segment TS is greater than the threshold τ.

After clustering, the cluster set Cl containing each cluster can be obtained. firstly, the cosine similarity matrix of each cluster is calculated; secondly, the similarity index set is calculated according to definitions 1-2; after that, according to definitions 1-3, the anomalous trajectory fragment can be obtained; finally, the anomalous trajectory is finally determined by definitions 1-4.

III. Experiment and Analysis

A. Experimental Data Set

The trajectory data set used in this paper is the trajectory data of one month (2018-12-19 00:00:00 to 2019-01-18 23:59:59) of the internal transport vehicle operations obtained from a port in Chongqing. Using the vehicle-mounted mobile terminal, data were collected with a sampling frequency of 8s/time, and a total of 1,239 operation instructions, 7,547 trips, and 3,415,835 trajectory points were collected.

B. Results and analysis

The parameters involved in the trajectory anomaly detection algorithm include open angle threshold ω, learning rate, nearest neighbor threshold ε, minimum number threshold MinPts, anomaly index threshold ρ and anomaly trajectory segment length ratio threshold τ. The settings of these parameters directly affect the results of trajectory anomaly detection.

The setting of the learning rate directly affects the optimization effect of the loss function in the self-encoder. The six curves from loss1 to loss6 are shown in Fig.4, which correspond to the loss functions with learning rates of 0.05, 0.1, 0.3, 0.5, 0.7 and 0.9, respectively.

The settings of the nearest neighbor threshold ε and the minimum number of MinPts have a direct impact on the clustering. However, the OPTIC algorithm is not sensitive to the parameters, and the reference value of the nearest neighbor threshold ε is given in the clustering, which greatly reduces the difficulty of parameter adjustment. After engineering experience and several experiments, the minimum number threshold MinPts is in the range of 15-25, and the nearest neighbor

threshold ε is in the range of 0.5-0.6, which is more suitable.

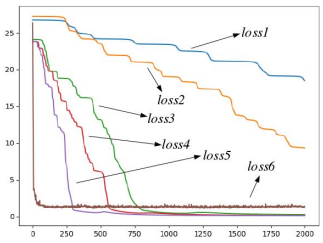


Fig. 4. Loss Function of Learning Rate.

Since the workload of manually calculating the result set is very large, the experiments in this section randomly select the trajectory data under 50 (45 ship unloading instructions and 5 transtacking instructions), 75 (67 ship unloading instructions and 8 transtacking instructions), 100 (90 ship unloading instructions and 10 transtacking instructions), and 125 (112 ship unloading instructions and 13 transtacking instructions) operation instructions in the data set, and after the extraction of trajectory real state feature sequences, the results are compared with the manually acquired result set, and the results are obtained as shown in Fig.5.

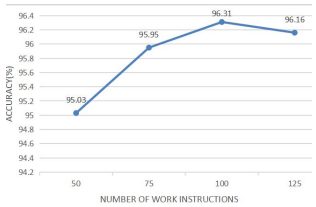


Fig. 5. Accuracy of Algorithms for Data of Different Sizes.

Overall, with the increase of the number of job instructions, the accuracy rate of S-DFFM is roughly at about 96% and the false alarm rate is only about 4%. The algorithms KNN(K-Nearest Neighbor), CNN(Convolutional Neural Networks), SVM(Support Vecor Machine), TSAD-FE(Trajectory Structure Anomaly Detection Based Feature Entropy), TAD-FD(Deep Characteristic) and the S-DFFM proposed in this paper are also selected, and the accuracy rate is averaged by several experiments to get as shown in Fig.6. Through the experimental results, it can be seen that the anomaly detection method based on unsupervised feature fusion proposed in this paper makes the similarity measure between trajectories more accurate and can prove the effectiveness of the algorithm and feasibility of the algorithm.

IV. Conclusion

In order to effectively supervise the cargo transfer process of port operation vehicles, this paper proposes a self-encoder-based feature fusion model (S-DFFM). The existing anomaly detection algorithm does not comprehensively consider the common features of the trajectory and is limited by the spatial

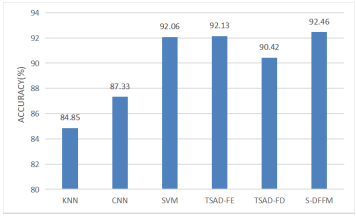


Fig. 6. Comparison of Accuracy Rates of Different Algorithms.

attributes of the trajectory, and this paper will fuse shallow feature sequences with deep feature sequences to represent the fused features of the trajectory, and the trajectory is represented by the feature sequences, which dilutes the spatial attribute of the trajectory, and the limitation of using spatial distance to judge the abnormality is solved. Based on the real operation trajectory data of a port in Chongqing, S-DFFM extracts the fusion features of the trajectory by using the fusion feature model and experiments the data set by using the unsupervised anomaly detection algorithm OPTIC. The experimental results show that the trajectory state extraction method of the fusion feature model in this paper can accurately extract the real state of the trajectory of the operating vehicle and can extract the stationary state of the vehicle during the operation. The experimental results show that the accuracy of S-DFFM in detecting anomalies in the trajectory of operating vehicles reaches 96%, which verifies that the trajectory anomaly detection algorithm proposed in this paper is suitable for detecting anomalies in the trajectory of operating vehicles in this port.

References

[1] Zheng Y. Trajectory data mining: an overview[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2015, 6(3): 29-71.
[2] Knorr E M, NgRT, Tucakov V. Distance-based outliers: algorithms and applications[J]. The VLDB Journal, 2000, 8(3): 237-253.
[3] Lee JG , Han JW, Li XL. Trajectory outlier detection: A partition-and-detect framework[C]. In Proceedings of the 24th IEEE Computer Society International Conference on Data Engineering, 2008: 140-149.
[4] Zhu J, Jiang W, Liu A, et al. Time-dependent popular routes based trajectory outlier detection[C]. In Proceedings of the 2015 International Conference on Web Information Systems Engineering, 2015: 16-30.
[5] Laxhammar R, Falkman G. Online learning and sequential anomaly detection in trajectories[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 36(6): 1158-1173.
[6] Laxhammar R, Falkman G. Sequential conformal anomaly detection in trajectories based on hausdorff distance[C].In Proceedings of the 14th International Conference on Information Fusion. IEEE, 2017: 1-8.
[7] Amodei D, Ananthanarayanan S, Anubhai R, et al. Deep speech 2: End-to-end speech recognition in english and mandarin[C]. In Proceedings of the 33rd International Conference on Machine Learning. 2016: 173-182.
[8] Ulyanov D, Vedaldi A, Lempitsky V. Deep image prior[C]. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9446-9454.
[9] AI-Saffar A A M, Tao H, Talab M A. Review of deep convolution neural network in image classification[C]. In Proceedings of the 2017 IEEE International Conference on Radar, Antenna, Microwave, Electronics and Telecommunications. 2018: 26-31.
[10] Dong W , Li J , Yao R , Li C, Yuan T, Wang L. Characterizing driving styles with deep learning[J]. arXiv preprint, 2016: 1607.3611.
[11] Chen C, Liao C, Xie X, Wang Y, Zhao J. Trip2Vec: a deep embedding approach for clustering and profiling taxi trip purposes[J]. Personal and Ubiquitous Computing, 2019, 23(1): 53-66.

*Research on Model and Method of Inland River Ship Overload Measurement**

Ai Lirong
Wuhan Institute of Shipbuilding Technology
Wuhan 430050, China
e-mail: 3247438@qq.com

Wang Yanchun
Huangshi Second High School
Huangshi, 435003, China
e-mail: 107370033@qq.com

Yi Jia, Jin Shengping
School of Science
Wuhan University of Technology
Wuhan 430070, China
e-mail: 1309059550@qq.com
spjin@whut.edu.cn

Abstract—The problem of detecting overloading of inland vessels is an important topic on water traffic safety. But the workload is too large to rely solely on traditional manual enforcement. This paper studies model and method for automatically calculating the distance between the trunk of an inland watercraft and the water surface by utilization of LIDAR(LIght/LAser Detect And Ranging) data and modern technologies such as AIS(Automatic identification System) and network transmission. Related models are established which calculate the real-time water surface height, preprocess the LIDAR data and measure the distance between the trunk and the water surface by a dual-fitting model. The model and method have their important theoretic meaning and can be applied to practice for the detection of overloaded inland watercraft.

Keywords- ship overload; laser radar; regression analysis; ship's actual freeboard value

I. INTRODUCTION

In recent years, the overloading of ships in the inland water network area has become more and more serious[1,2]. At the same time, accidents such as ship sinking, casualties, and channel obstruction have occurred from time to time. This affects not only the safety of water transportation, but also restricts the development of water transportation economy. The phenomenon of ship overloading has been repeatedly banned due to that it is a hidden danger to the safety of water traffic. As an important business task of maritime management agencies, the prevention and control of overloading has always been highly valued by maritime departments[12]. In the practice of preventing overloading, the detection of ship overloading was mainly carried out manually. With the increase of ship flow and the continuous improvement of modern management level, it is necessary to study the automatic measurement method of ship overloading. This problem involves the comprehensive integration of multiple equipment and technologies[13]. The key issue is the model and calculation method for measuring the height of the edge points on the deck surface and the water surface.

Some scholars have researched into the issue already. There are many methods such as sonar technique[6], video graphic[4], pressure sensor[9], electronic water gauge[10], DSP etc. These methods have a common shortcomes in which some sensors are required to be installed into ships[10,11]. The laser scanners are also used to check the height betheew the free board and the water surface. But the models are complicate in which the video graphic are needed also[7,8,14].

If this problem can be solved, the static and dynamic information of the ship can be automatically obtained through the relevant position information of the AIS[15] equipment, and the ship database of the maritime department can be accessed to obtain its registered dry height[3,5]. By comparing the two heights, It can be judged whether the ship is overloaded or not.

Fig. 1 shows the main equipment of the proposed ship overload detection system. The LIDAR continuously scans the ships passing through the detection area to obtain the cross-sectional data of the ship, and through the analysis and calculation of the obtained cross-sectional data of the ship, the actual height from the edge point of the upper surface of the freeboard deck to the water surface is obtained to judge whether the ship is overloaded.

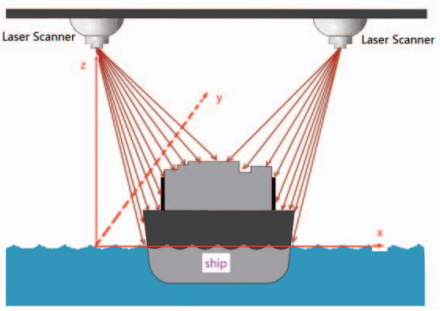


Figure 1. Ship overload detection system.

II. LIDAR EQUIPMENT SELECTION AND REAL-TIME CALCULATION OF WATER SURFACE HEIGHT

A. LIDAR related terms

LIDAR is a way to obtain directly three-dimensional coordinates of object surface points through observation data such as position, distance, and angle. It can be applied to a series of application scenarios such as ship surface information extraction and three-dimensional scene reconstruction technology. The LIDAR data involved in this article is two-dimensional cross-sectional data, which is to obtain the angle and distance information of the ship's surface under a certain angular resolution.

1) Resolution: The minimum change value of the detection distance that can be output. To change the default, adjust the template as follows.

*It is Supported by: Technological innovation team work on ship's new energy and power equipment application, No.2019td01

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00024

2) Angle resolution: The minimum change value of the detection angle that can be output, in radians.

3) Horizontal resolution and vertical resolution: The minimum change value of the distance in the horizontal direction and the vertical direction. To change the default, adjust the template as follows.

4) Systematic error: A series of observations are made under the same observation conditions, and the magnitude and sign of the error show a fixed pattern or change according to a certain law.

5) Random error: Refers to the deviation between the measured value and the true value random.

6) Free board (sometimes referred to as standard free board): Refers to the vertical distance measured from the upper surface of the free board deck to the relevant full-load waterline in the middle of the ship. The calculation formula is

$$F = D + \delta - d(m).$$

In this formula: F is free board(m), D is moulded depth(m), δ is the thickness of freeboard deck side plate(m), and d is ship's full draught(m).

7) The distance between the free board deck and the surface of the water: Refers to the average distance from the upper surface of the free board deck to the surface of the water in the middle of the ship(m).

B. Real-time calculation of water surface height

The height of the water surface of the inland river is changing continuously due to the influence factors such as tides, rain, fluctuations, wind and waves, so it is necessary to calculate the height of the water surface in real time.

As shown in Fig. 2, a dedicated LIDAR is used to measure the real-time height of the water surface.

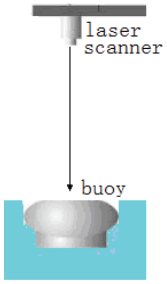


Figure 2. LIDAR measures the height of the water surface in real time.

h_0 is the previously calculated height of the water surface from the radar before calculating the altitude, and the value of h_0 for the first time can be determined by actual measurement.

Suppose there are a total of several cross-section data $\{(\rho_i, \theta_i) | i = 1,2,3\}$, we select three points in the k-th cross-section data, where θ_i is the angle between the reflection line and the straight downward line. They meet the following conditions: the value of $|\theta_i|$ is the smallest 3 of the cross-section points; $\theta_1 < \theta_2 < \theta_3$. Let $z_i = \rho_i \cos\theta_i, i = 1,2,3$, under ideal conditions, the average height of the water surface from the radar calculated from this cross-section is $h_k = \frac{1}{3}\sum_{i=1}^{3} z_i$.

The $k-th$ section data should be eliminated if

a) $\max\{|z_1 - z_2|, |z_2 - z_3|, |z_1 - z_3|\}$ is obviously too big.

b) The h_k calculated by the (1) such that $|h_k - h_0|$ obviously too large.

c) $z_1 + z_3 = 2z_2$ is obviously not true.

Suppose there are m remaining sections after the elimination. Then the height h_0 of the water surface from the radar is updated to the real-time height.

$$h_0 = \frac{1}{m}\sum_{k=1}^{m} h_k \tag{1}$$

The h_0 calculated by (1) is saved as a system parameter and recorded as the initial value for the next real-time calculation of altitude.

III. THEORETICAL MODEL AND CALCULATION METHOD

For the convenience of discussion, we first establish a right-handed spatial coordinate system, as shown in Fig. 1. The line connecting the two LIDARs installed under the bridge deck is the x-axis direction, the upward direction perpendicular to the water surface is the z-axis, and the lower displacement h_0 of the LIDAR is the coordinate origin. It is the vertical height between the LIDAR and the horizontal water surface, and its value is obtained by the LIDAR from a buoy placed on the water. The coordinate plane xoz is the plane where the laser radar scans to obtain the ship's cross-sectional data. The angle of the laser radar scan is based on the negative direction of the z-axis, which is adjusted when the laser radar is installed. The y-axis direction is the direction the ship travels. In other situations, such as the other direction of the ship, the LIDAR on the other side, etc., the modeling conditions can be transformed into this situation by simply rotating and translating the coordinate axis. In order to facilitate data exchange with AIS and GPS, the longitude and latitude of the LIDAR and the direction of the y-axis in the GPS coordinate system should also be determined.

A. Data conversion and preprocessing of LIDAR

At a certain moment when the LIDAR scans the ship, it scans the outer surface of a certain section of the ship, and obtains the reflection data of a series of points, namely a series of distances and angles $P_i(\rho_i, \theta_i), i = 1,2,\cdots,n$, which is shown in Fig. 3. $OR = h_0$ is the height of the LIDAR from the water surface, $\angle ORP_i = \theta_i$ is the angle, and the angle direction is defined as the positive direction starting from RO and counterclockwise with R as the center. The rectangular coordinates of P_i are:

$$x_i = \rho_i \sin\theta_i, z_i = h_0 - \rho_i \cos\theta_i \tag{2}$$

Because the data is noisy, the data points after (2) conversion need to be preprocessed, which is described as following 3 steps.

Step1. Floating objects on the water, sand and waves in the river water are in the rectangular coordinate system xoz, and the z coordinate must be near 0, so the point $z_i < \Delta_1$ in $P_i(x_i, z_i), n = 1, 2, \cdots, n$ is first eliminated. Where Δ_1 is a certain given threshold.

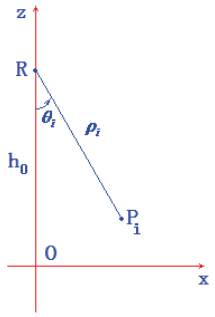


Figure 3. rectangular coordinate system for the cross section.

Step2. The possible reflection points of floating objects in the air, rain, snow, fog, etc. are in the rectangular coordinate system xoz, and the coordinate must be very large, so the point $z_i > \Delta_2$ in $P_i(x_i, z_i), n = 1, 2, \cdots, n$ is eliminated. Where Δ_2 is a threshold setted manually.

Step3. After the above two steps, the remaining data are arranged in ascending coordinates according to the x coordinates, and the one that $\frac{1}{3}$ is ranked first is a cross-sectional data point to be further processed.

B. Calculation model and method

1) Data refinement of cross-sectional feature points

Let $(x_i, y_i, z_i), i = 1, 2, \cdots, n$ be the n points extracted preliminarily by the cross-section feature point extraction model based on the adaptive threshold. The specific steps are as follows:

a) Calculate the sample mean and sample standard deviation of height $z_i\ (i = 1, 2, \cdots, n)$:

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i, s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2} \quad (3)$$

If $|z_i - \bar{z}| > 3s$, then delete the i-th point (x_i, y_i, z_i).

b) If the i-th point is deleted, go back to the i-th section data and try to find a point that is close to the $(i-1)$-th feature point $(x_{i-1}, y_{i-1}, z_{i-1})$ and the $(i+1)$-th feature point $(x_{i+1}, y_{i+1}, z_{i+1})$. The similar criterion is

$$\left| x_{ij} - \frac{x_{i-1} + x_{i+1}}{2} \right| < \delta, \left| z_{ij} - \frac{z_{i-1} + z_{i+1}}{2} \right| < 3s.$$

Where (x_{ij}, z_{ij}) is a certain point in the i-th cross-sectional data, threshold δ is the horizontal resolution corresponding to point $\left(\frac{x_{i-1} + x_{i+1}}{2}, \frac{z_{i-1} + z_{i+1}}{2} \right)$, and s is the sample standard deviation calculated by (3).

c) Repeat a) and b) several times until s is relatively small.

2) "Double fitting-removal" model of free board height

The y coordinate of the i-th cross-section feature point is represented by y_i, which is obtained by the speed v of the ship and the interval t_0 between the two scanning sections of the LIDAR:

$$y_i = v(n-i)t_0$$

Where n is the total number of all initial cross-sections of the LIDAR, and the y coordinate of the n-th cross-section is 0. The influence of the direction change of the speed v on the y coordinate and the influence on the feature point extraction algorithm can be ignored.

The ship's driving direction and its changes can be obtained through the ship's AIS real-time dynamic data. The calculation method is as follows: (1)Linearly interpolate the heading angle of the AIS ship every 1 second; (2)Numerical difference of direction angle; (3) Take the average of the calculated rate of turn at intervals (such as 10 seconds).

If the absolute value of the steering angle exceeds a certain range, the feature point of the corresponding cross-section is selected as a segment point during curve fitting. Each segment is fitted with a parabola of x and y.

If the absolute value of the steering angle does not exceed a certain range, use a straight line parallel to the y axis to fit.

The specific steps of the regression analysis from the cross-section feature point to the upper-surface feature points of the deck are composed of two repeated "fitting-removing":

Let $(x_i, y_i, z_i)(i = 1, 2, \cdots, m_1)$ be the sequence obtained by preprocessing the cross-section feature point data, generally $m_1 < n$.

- *Fitting-elimination process of x with respect to y*

a) Fitting m_1 points (x_i, y_i), and the fitted equation is recorded as

$$x = f(y) \quad (4)$$

If $|x_i - f(y_i)| > \delta$, delete the i-th point (x_i, y_i, z_i) in the m_1 data sequences. Where δ can be taken as $\frac{4}{3}$ times the horizontal resolution at (x_i, z_i).

b) When deleting data using (4) for the first time, if

there are a large number of points to be deleted in the first small part or the last small part of data points, we need to delete the entire section to eliminate the influence of the irregular curve formed by the edge points on the upper surface of the bow or stern deck.

c) Repeat *a),b)* several times for the remaining data points

- *Fitting-elimination process of z with respect to y*

Suppose that after the fitting-elimination process of x with respect to y, the remaining data point sequence is $(x_i, y_i, z_i)(i = 1,2,\cdots,m_2)$. Note that m_2 here is the new number of data points after the abnormal points have been eliminated.

a) Perform straight line fitting on m_2 points (x_i, z_i), and the fitted equation is recorded as $z = cy + d$. If $|z_i - cy_i - d| > \delta$, delete the i-th point (x_i, y_i, z_i) in the m_2 data sequence. Here δ can be taken as 3 times the mean square error of the sample points about z.

b) After two repeated "double fitting-removal" processes, m data points $(x_i, y_i, z_i)(i = 1,2,\cdots,m)$ are finally obtained. Pay attention to keep the correspondence between these m data points and the original original cross-section labels of the LIDAR relationship. Through the final fitting equation $z = cy + d$, the final estimated value $\bar{z}$ of the distance between the ship's deck surface and the water surface can be obtained:

$$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i, \bar{z} = c\bar{y} + d$$

IV. CONCLUSIONS AND SUGGESTIONS

This article mainly studies the selection of LIDAR and equipment installation, real-time calculation of water surface height, adaptive cross-section feature point extraction, cross-sectional feature point data preprocessing, cross-sectional feature point extraction to the upper surface feature point of the deck edge, deck on both sides of the shape and verification of the projection curve of the feature points on the edge of the upper surface, and the final method of determining the free board height, etc.

The data processing and modeling method of the single LIDAR above will be further extended to the comprehensive processing of two LIDAR, and the analysis and calculation on each side of xoz plane, xoy plane and yoz plane will finally get the height required to measure.

Specific engineering applications also involve the selection and installation of laser radar equipment, especially the alignment of the scanning plane, the alignment of the scanning angle, and the determination of the position and direction of the laser radar equipment in the GPS coordinate system. At the same time, the comprehensive integration of related equipment such as LIDAR, AIS, and cameras, its time synchronization, and the acquisition and utilization of static and dynamic data of ships, are also tasks that need to be further clarified.

In addition, the successful application of the entire system will also involve network transmission, database technology and computer software and hardware platform selection and other related issues. According to the model and method in this article and various hardware integration methods, implementation development and engineering applications can be carried out smoothly.

REFERENCES

[1] W. Chen, J. Yu, J. Xu, C. Jiang and L. Chen. A new measurement system of ship draft. Shipbuilding of China, 2013, 28(1), pp.166-171.

[2] C. David. Method and System for Surveillance of Vessels[P]. United States Patent: 20060244826A1, 2006-11-02.

[3] M. D. Graziano, M. D'Errico, and E. Razzano. Constellation analysis of an integrated AIS/remote sensing spaceborne system for ship detection. Advances in space research, 2012, 50(3), pp.351-362.

[4] T. Gu. Design of ship draft detection system. Nanjing University of Technology. 2012

[5] M. Hidenari, M. Furusho, and Y. Yano (2012). Analysis of ship evacuation during tsunami using AIS (Automatic Identification System) data. Journal of Earth Science and Engineering, 2012, 2(7), 412.

[6] Q. Huang. Design of Ship Overload Detection System Based on Multi Beam Sonar. Ship Electronic Engineering, 2016,036(010): 126-131.

[7] C. Mertz, L. Navarro-Serment E., R. MacLachlan, P. Rybski, A. Steinfeld, A. Suppe, J. Gowdy. Moving object detection with laser scanners. Journal of Field Robotics,2013, 30(1), pp.17-43.

[8] K. Nakamura, H. Zhao, R. Shibasaki, K. Sakamoto, T. Ohga, and N. Suzukawa. Tracking pedestrians using multiple single-row laser range scanners and its reliability evaluation. Systems and computers in Japan, 2006,37(7), pp.1-11.

[9] X. Ran, C. J. Shi., J. Chen, S. Ying, and K. Guan. Draft line detection based on image processing for ship draft survey. In Proceedings of the 2011 2nd International Congress on Computer Applications and Computational Science. Springer, Berlin, Heidelberg. 2012, pp. 39-44.

[10] H. Su. Study on dynamic measurement method of ship draught. Wuhan University of Technology. 2008.

[11] M. Xiong, S. Zhu, L Li, and P. Ni. Research on data processing method of real-time detection system for dynamic ship draft. Chinese Journal of Scientific Instrument, 2012,33(1), pp.173-180.

[12] X. Yan. Design of the ship's draft depth detection system. Ship Electronic Engineering,2016, 038(04): pp.49-51.

[13] G.Yang, X. Yang, T. Xiong. Electric immergence monitoring system for watercraft: CN, CN100537350 C[P]. 2009.

[14] Y. Zhao, G. Lu, X. Guo, and Y. Wang (2014). Vessel Freeboard Calculation Method Based on Laser Scanning. In Intelligent Data analysis and its Applications, Volume I. Springer, Cham. 2014, pp.299-307.

[15] S. Zhu, Z. Liu, W. Jiang, and K. Guo. The Key technology of Blind Source Separation of Satellite-Based AIS. Procedia Engineering, 2012,29, pp.3737-3741.

Research on Path Planning Algorithm of Intelligent Wheeled Robot Based on ROS

Wang Ming-yuan
College of Information &Control Engineering
Jilin Institute of Chemical Technology
Jilin, China

Zhao Meng-yao
College of Information &Control Engineering
Jilin Institute of Chemical Technology
Jilin, China

Zhu Jian-jun
College of Information &Control Engineering
Jilin Institute of Chemical Technology
Jilin, China
e-mail: zjj099@163.com

***Abstract*—Aiming at some limitations of the A-star algorithm, the algorithm is improved and optimized, robot path planning experiment is carried out. In the traditional A-star algorithm, there are too many turning points of the path and the path is not smooth enough, the car deviates from the path when navigating due to the arc at the corner is too large. This paper combines the conjugate gradient and Gaussian process principle to optimize the A-star algorithm and eliminate the invalid path information to jump points, to implement a path with less angle and smoother. The results show that the hybrid algorithm has a shorter and smoother path than the A-star algorithm.**

Keywords-component; A-star algorithm; path planning; conjugate gradient; Gaussian process; jumping point

I. INTRODUCTION

With the rapid development of technology, intelligent devices are gradually entering people's daily lives. Smart wheeled mobile robots are a major branch of smart applications and have a wide application in daily life, such as industrial applications, medical catering and other industries. The autonomous navigation of robots is the most important part of intelligent wheeled mobile robots [1], it requires mobile robot to plan a path based on some optimization criteria or indicators (such as the smallest energy consumption, the shortest completion time, the smoothest path, the shortest walking path, etc.) The mobile robot is required to plan a safe and collision-free shortest trajectory from the start point to the target point [2]. Due to the expansion of the working range of mobile robots, the requirements of path planning technology are also increasing. Currently, the most widely used path planning algorithms include the A-star algorithm, ant colony algorithm, DWA (Dynamic Window Approach) algorithm, artificial potential field method, Dijkstra algorithm, etc. [3-6] Path planning is one of the key factors in autonomous robot navigation. The advantage of the A-star algorithm is that it can find the best path, but in a complex environment, the A-star algorithm takes a longer time to find the path and the path is not reach the optimal path. The path planned by the A-star algorithm is an angular turn at the corner and the arc-like turn required by the car is not considered.

Literature [7] proposed an improved ant colony algorithm, this algorithm uses gradient descent to optimize the path of the ant colony algorithm, which improves the convergence speed and accuracy of the ant colony algorithm. Literature [8] adopts the A-Star algorithm based on changing the step size to reduce the calculation time of the algorithm and improve the stability and robustness of the algorithm performance. Literature [9] uses a three-domain search A-star algorithm combined with artificial potential fields to optimize the path planning problem of mobile robots, which shortens the length of the path, reduces the search time and the number of nodes. Literature [10] uses the hybrid-A star algorithm, which is optimized by the conjugate gradient method and is suitable for Ackerman vehicles. Because kinematics is added to plan the vehicle's backward route, the overall planned route may be longer than the route planned by the A-star algorithm, but this algorithm improves the smoothness and safety of the route. Literature [11] proposed a path planning algorithm based on jumping point search combined with the Bezier curve to achieve path optimization.

II. AN IMPROVED HYBRID ALGORITHM BASED ON A-STAR ALGORITHM

The A-star algorithm estimates the optimal path based on the step length and cost. The JPS (Jump Point Search) algorithm speeds up the operating efficiency of the A-star algorithm by reducing the number of nodes traversed, reducing the length of the path and the number of corners turned. However, the JPS algorithm also has problems such as the path is not smooth enough and the corners are too large. The algorithm is optimized by combining the advantages of the JPS algorithm, first, perform conjugate gradient processing on the A-star algorithm to obtain a piece of new path information. Then the path information is pruned, only part of the optimal path information is retained, and then the Gaussian process is performed on the path to obtaining an optimal and smooth path.

The principle of the jumping point is to improve the heuristic function of the A-star algorithm, the expression is as (1).

$$h(n)=\lambda L - k\cos a \quad (1)$$

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00025

L is the Manhattan distance from the start point to the target point, the expression is shown in (2), λ and k are the weights of distance and direction, cos$a$ is the direction information of the angle between the parent node and the child node, and the child node and the target point. When the two directions are the same, the value of cos$a$ is + 1, and the direction cos$a$ is the smallest, and vice versa, the value of cos$a$ is -1, the direction has the largest cost.

$$L = \beta \ (\text{abs}\ (n.x - \text{goal}.x) + \text{abs}\ (n.y - \text{goal}.y)) \quad (2)$$

β is the weight, n.x and n.y are the abscissa and ordinate of the current position node, goal.x and goal.y are the abscissa and ordinate of the target node respectively.

The Gaussian process is a commonly used optimization method, which is widely used in image processing and curve processing. The Gaussian process is often used in machine learning, where the core of the Gaussian process is the kernel function. The final results obtained by applying different kernel functions will also be different. The measurement method of each kernel function is different and the processing properties of the Gaussian process are also different. The more commonly used kernel function is the Gaussian kernel function, which is also called the radial basis function (RBF), the expression is shown in (3).

$$K(x,x') = \delta^2 \exp\left(-\frac{(x-x')^2}{2l^2}\right) \quad (3)$$

Where x and x' are the abscissa of the sampling points, and δ and l are the hyperparameters of the Gaussian kernel. Figure 1 uses sin$x$ as a sample function, where the green line is the prediction curve, the red points are the sampling points and the light blue is the variance confidence interval. When l=0.5, δ=0.5, the curve has more corners and more twists and turns compared to the other three pictures. When l=0.5, δ=1.0, the curve is smoother than that in Figure a, but it is still not smooth enough. When l=1.0, δ=0.5, the curve becomes smoother but the smoothness is higher. When l=1.0, δ=1.0, the smoothness and gentleness of the curve are relatively the best. From the four graphs of a, b, c, d, it can be concluded that the l increasing curve is smoother and the confidence interval between the sampling points is smaller. Otherwise, the curve is more tortuous. The value of δ can control the size of the confidence interval. When the value of δ increases, the confidence interval increases, and vice versa.

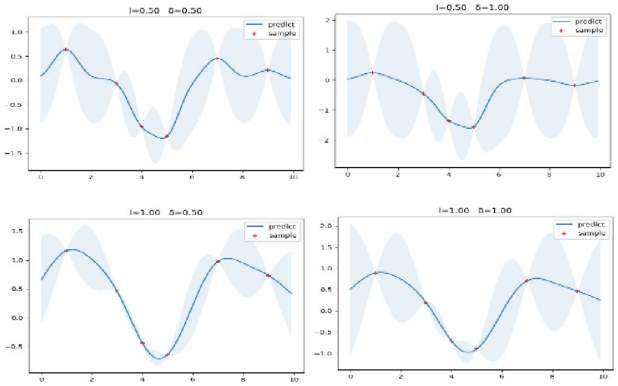


Figure 1. Gaussian process curve

Find the optimal parameters δ and l by maximizing the edge likelihood, the formula is shown in (4).

$$p(y|\delta,l) = \frac{1}{(\sqrt{2\pi})^n \sqrt{K_{yy}}} e^{-\frac{1}{2}y^T K_{yy}^{-1} y} \quad (4)$$

After logarithmic processing, formula (5) is obtained.

$$\log p(y|\delta,l) = -\frac{1}{2}y^T K_{yy}^{-1} y - \frac{1}{2}\log|K_{yy}| - \frac{n}{2}\log(2\pi) \quad (5)$$

Maximizing the logarithmic marginal likelihood refers to the maximum marginalization on the function value y, where K_{yy} is the covariance matrix. After derivation, the optimal hyperparameters are l=1.2, δ=0.8, and the test results are shown in Figure 2 show.

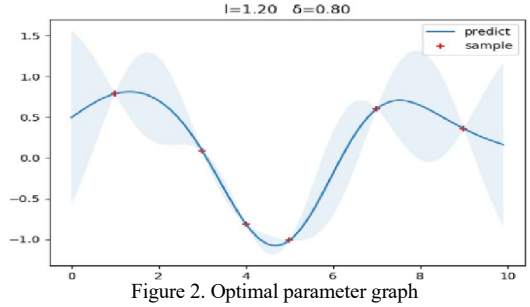


Figure 2. Optimal parameter graph

The flowchart of the improved hybrid algorithm is shown in Figure 3. The starting point and the target point are set, the path is planned by the conjugate gradient A-star algorithm, and the path is skipped. Gaussian smoothing is performed after the path that eliminates interference is obtained, and the optimal smooth path is obtained.

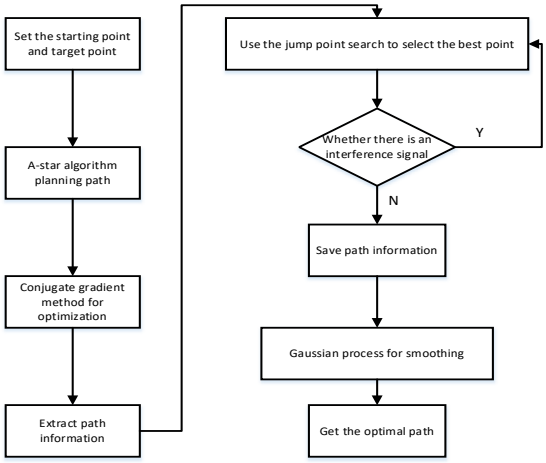


Figure 3. Hybrid algorithm flow

III. SIMULATION EXPERIMENT COMPARISON

A. *Without obstacle simulation*

Optimize the path without obstacles and test it on the maps of 3x3 and 9x14 specifications. The red line is the unprocessed path and the blue line is the optimized path, the test results are shown in Figure 4 and Table 1.

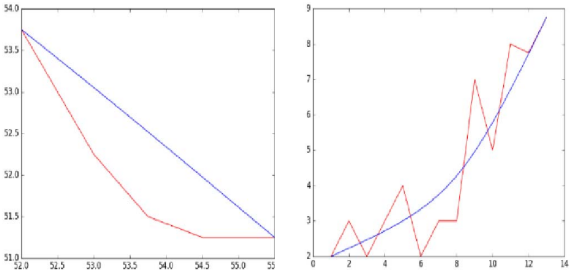


Figure 4. Without obstacle simulation

TABLE 1. Simulation data

Map Specifications	Path Length (cm)	
	unprocessed path	*optimized path*
3x3	5.201	4.301
9x14	26.733	19.065

The data comparison shows that the processed path is shorter than the unprocessed path and the path is smoother.

B. With obstacle simulation

This simulation experiment is set up in Gazebo as shown in Figure 5, the planned paths are displayed in Rviz. Figure 6 shows the A-star, Conjugate gradient, Jump point search and Gaussian smoothed path from top to bottom. The path after the conjugate gradient processing is smoother than the path of the A-star, but the overall path is not smooth enough. After the jumping point processing, the invalid path information is eliminated and the path between the two turning points is straighter. Then the Gaussian process optimization is performed on the path after the jump point, and a smooth optimal path is obtained.

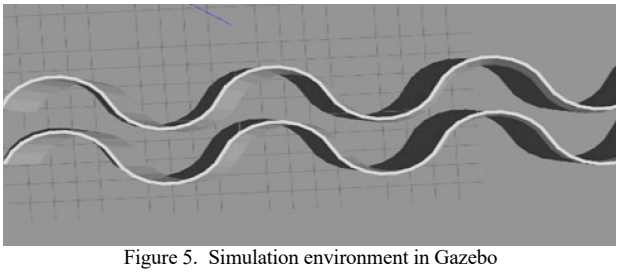


Figure 5. Simulation environment in Gazebo

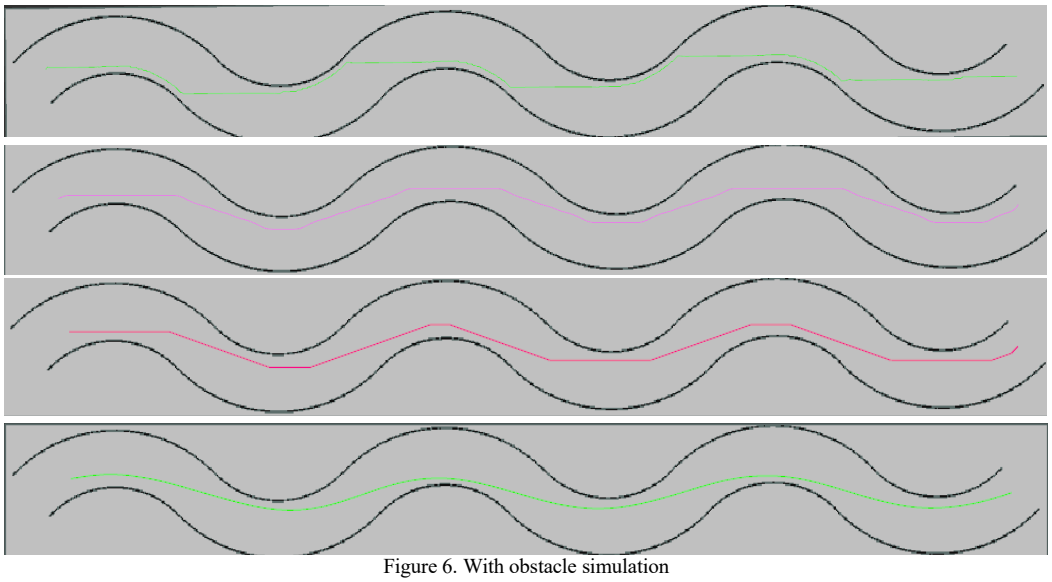


Figure 6. With obstacle simulation

TABLE 2. Simulation data

Simulation Map	Race Track			
	A-star	*Conjugate gradient*	*jump point*	*Gaussian smoothed*
Path length	81.254	77.782	75.554	70.251
Number of nodes	7322	5502	573	573

By comparing the path length with the number of points traversed, the A-star algorithm path has more nodes and longer paths. The number of nodes in the path after conjugate gradient processing decreases and the path length is shorter than the path of the A-star algorithm. The jumping point is the selection of path nodes on the path after gradient processing. The number of nodes traversed by the path decreases significantly, and the path length is slightly shortened. The length of the path smoothed by the Gaussian process is significantly shortened, and the number of nodes is equal to the number of nodes processed by the hop point. The optimized hybrid algorithm preliminarily solves the non-optimal path of the A-star algorithm in a complex environment, the path is not smooth enough, the corners are too much.

IV. NAVIGATION SIMULATION EXPERIMENT

In order to verify whether the path planned by the hybrid algorithm meets the motion requirements of the mobile robot in a complex environment, a path navigation simulation experiment is carried out for the mobile robot, the robot runs along the planned path from the starting point to the specified target point. Figure 7 shows the simulation results of robot navigation, where the green line is the route planned by the hybrid algorithm and the red line is the historical trajectory of robot operation. It can be seen from the figure that the route of robot operation is consistent with the route planned by the algorithm, which verifies the feasibility of the algorithm.

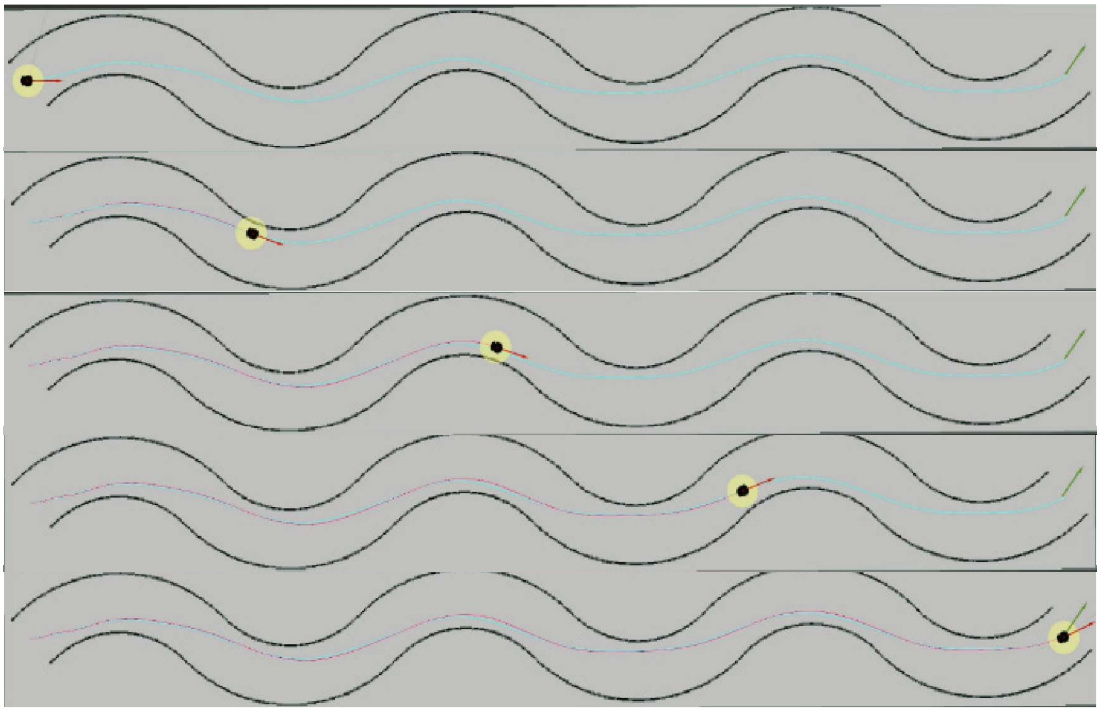


Figure 7. Navigation experiment simulation

V. CONCLUSION

Experiments show that the path planned by the improved hybrid algorithm in this paper is better than the path planned by the A-star algorithm. It reduces the number of corners and the arc of the mobile robot in a complex environment, the path is smoother. In the robot navigation simulation experiments, the robot's forward trajectory is the same as the planned route, which provides an effective algorithm for the mobile robot to quickly and stably run to the target point.

REFERENCES

[1] Zhang Xin, Zhang Yu, Su Xiaoming.Design and Implementation of Autonomous Positioning and Navigation System for Mobile Robots, Machine Tool and Hydraulics, vol 48, 2020, pp.88-91.

[2] Zhang Jinyue, Hou Zhicheng, Zhang Gong, et al.Improved A* Algorithm Based on Traffic Constraints and Multivariate Heuristic Function. Modular Machine Tool and Automatic Processing Technology, 2021(01):53-56.

[3] Dai Xiaolin, Long Shuai, Zhang Zhiwen, Gong Dawei. Mobile Robot Path Planning Based on Ant Colony Algorithm With A* Heuristic Method. Frontiers in neurorobotics, vol. 13, 2019, pp.15.

[4] Qin Yuannian, Liang Zhonghua.New progress in research and application of ant colony algorithm.Computer Engineering and Science, vol.41, 2019, pp.173-184.

[5] Jiang Weinan. Research and Application of AGV Path Planning and Autonomous Obstacle Avoidance[D]. Changchun: Changchun University of Technology, 2020.

[6] Dong Xuanliang, Zhao Guiqing.Robot navigation path planning based on artificial potential field-guided ant colony algorithm.Machinery Design and Manufacturing, 2021(06):169-173.

[7] Qin Kun, Guan Zequn, Li Deren, et al.Research on the best path analysis method based on raster data.Remote Sensing for Land and Resources, 2002(02):38-41.

[8] Shang Erke, Dai Bin, Nie Yiming, Zhu Qi, et al. An improved A-star based path planning algorithm for autonomous land vehicles. International Journal of Advanced Robotic Systems, vol.17, 2020, pp.1729881420564952.

[9] Chen Jiqing, Tan Chenzhi, Mo Rongxian, et al. Research on path planning of three-neighbor search A* algorithm combined with artificial potential field. International Journal of Advanced Robotic Systems, vol.18, 2021, pp.17298814211026449D.

[10] Dušan Nemec, Michal Gregor, Emília Bubeníková, et al. Improving the Hybrid A* method for a non-holonomic wheeled robot. International Journal of Advanced Robotic Systems, vol.16, 2019, pp.1729881419826857.

[11] Zhang Ben, Zhu Denglin. A new method on motion planning for mobile robots using jump point search and Bezier curves. International Journal of Advanced Robotic Systems, vol.18, 2021, pp.32.

Research on scheduling algorithm of multi-AGVS system

Yao Xiang
Wuxi Taihu University, Wuxi 214064, China
Jiangsu Key Lab of IoT Application Technology,
Wuxi Taihu University,Wuxi, China 214064

PENG Li
Wuxi Taihu University, Wuxi 214064, China
Jiangsu Key Lab of IoT Application Technology,
Wuxi Taihu University,Wuxi, China 214064

Abstract—**Due to the rapid development of global economy, the demand for Automated Guided Vehicles System(AGVS) scheduling is increasing in the production and transportation industry.As an automatic guide device that can transport a specified item to a specified location, Automated Guided Vehicles (AGV) greatly reduces the transportation efficiency and transportation cost.Among the basic AGVS scheduling, task management and path control are the two most widely studied and valued tasks.We construct a new integration model for scheduling problem and conflict-free routing problem of multiple AGVs.A new mathematical programming model is developed, and the ant colony algorithm (ACA) is applied to solve the model.The algorithm is optimized based on multi-objective programming, working similarity and pheromone matrix.A scheduling experiment is used to find the optimal path, and the performance of ACA is compared with that of precise algorithm.The results show that although the precise algorithm can accurately solve small scale problems, the improved ant colony algorithm is more suitable for medium and large scale problems than the precise algorithm, which shows the credibility of ACA to solve medium and large scale problems.Therefore, it can be concluded that the improved ant colony algorithm has the ability to provide task management and path planning for large-scale AGVS scheduling problems in an acceptable time range.**

Keywords: AGVS;Scheduling algorithm;Ant colony algorithm

I. Introduction

In the process of warehousing and logistics, loading, unloading, transportation and other processes take up most of the time of the whole logistics process. AGV is recognized as a good solution for material handling because of its higher efficiency and good flexibility, which reduces the material transportation cost in the total cost and improves the space utilization rate of logistics warehouse[2].In these indoor logistics systems, it may be necessary to use multiple AGVs to perform multiple transport tasks at the same time.Therefore, these AGVS are critical to planning an efficient AGV system (AGVS) that is conflict-free and collision-free.AGVS is the general name of a whole set of systems composed of automatic guided vehicle, upper control system, guidance system, communication system and charging (supply) system[1].However, the design and control of AGV system is not easy to operate, and the biggest challenge comes from the scheduling system for task management, vehicle monitoring and traffic management in the upper control system[1].Therefore, the establishment of an efficient multi-AGVS scheduling system can further improve the structure of logistics and transportation system, reduce logistics and transportation costs, improve system operation efficiency, and has important application value in intelligent warehousing and logistics.However, in the real warehouse, due to the dynamic changes of AGV storage location and environmental information, complex road conditions may lead to congestion and collision problems. Therefore, screening the optimal task set and path for AGV is the most fundamental task.

The production shop considered in this paper is arranged with the same embedded dedicated network guide rails which do not allow AGV to move in all directions, i.e. AGV can only move along the guide rails.In the early stages of conflict-free routing scheduling, basic information such as the number and location of machines, the process of the job to be processed, and the number of vehicles is first understood[4]. The ultimate goal is to find the correct processing sequence in which the workpiece is processed. Ensure conflict-free transportation routes and process all parts in a minimum time frame [5].

II. Improvement and simulation of ant colony algorithm

A Ant colony algorithm overview

Ant colony algorithm (ACA) is a famous meta-heuristic intelligent search algorithm first developed by Dorigo, Maniezzo and Colorni (1991) in the early 1990s[9], the best path can be found through distributed collaboration. See Figure 1,a swarm of ants is moving along A and E.Suppose A is the ant's nest and E is the food the ant finds.The ants are going to move along the AE line.But if a rock suddenly appears between A and E, the ant at B or D must decide whether to drive left or right.Since the ant in front didn't leave any pheromones behind, the ant could have moved in either direction.But when an ant passes by, it releases pheromones in its path, one of the ants' tools for communicating with each other.Pheromones disappear at a certain rate.The ant behind it decides whether to go left or right based on the pheromone concentration in the path.It is clear that pheromones will become increasingly concentrated on the short path, attracting more and more ants to move along the path.

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00026

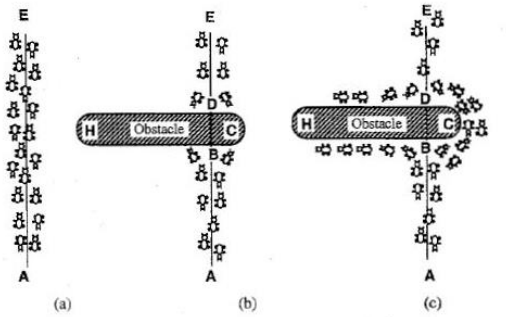


Figure 1 Schematic diagram of ant colony algorithm

B The advantages of ant colony algorithm

Ant colony algorithm as a multi - agent random search algorithm[12], has a strong optimization ability, is widely used in the workshop scheduling, vehicle routing and other problems.However, ant colony algorithm also has some unique disadvantages[8].(1)The operation time of the ant colony algorithm is relatively long.In the face of more iterations and ant numbers, the search speed of the algorithm will become slower.(2) The feedback mechanism of ant colony algorithm is a positive feedback mechanism. Although the algorithm has a good convergence speed, it is easy to fall into the local optimal solution and difficult to jump out[7].(3) Although ant colony algorithm is suitable for solving various problems, it has the problem of optimization ability.For example, when solving path planning problems, taboos set on each node will cause the program to appear "deadlock".[10].

C Algorithm improvement

In the research of ant colony algorithm at home and abroad, there are many improved algorithms, such as the improvement based on pheromone, based on the improvement of heuristic information, this paper mainly focuses on the adaptive ant colony algorithm for discrete domain optimization problems.

- *To improve the direction*

In the ant colony algorithm, the heuristic information is improved[3].The multi-AGVs scheduling problem studied in this article is as follows: In the shop environment of an operating system, one or more machines are required to handle many jobs.These jobs take different paths and require different machines to complete the process.Therefore, this paper makes the following assumptions: (1) All AGVs have unit working capacity and will not fail during continuous operation;(2) All AGVs are in standby state, all jobs are in the state of waiting to be processed, and no new jobs can be added after the processing sequence is determined;(3) The starting point and ending point are on the network;(4) AGV loading time is ignored;(5) In the next time unit, only one job is assigned to each AGV, and no machine can handle more than one job at a time;(6)AGV moves operations from the starting point of the warehouse to the first pick up and delivery point of the required machine via a network-boot path;(7)The vehicles carrying the job cannot transfer the job to other vehicles, they must wait for the machine to complete the job processing, and if necessary receive the job and transfer it to the subsequent machine.Otherwise, they move to the end of the warehouse;(8)AGVs should not collide with each other in the course of routing, i.e., to achieve conflict-free routing.

- *Algorithm to improve*

(1) Global update rules

In the general ant colony algorithm, a transfer update is performed after the ant colony finishes a path search. However, in the improved formula 1, the update is only applicable to the ants with the optimal path search, that is, only the pheromone content on the path of the global optimal solution is increased.

$$\begin{cases} \tau_{ij} = (1-\rho)\cdot\tau_{ij} + \rho\Delta\tau_{ij} \\ \Delta\tau_{ij} = \begin{cases} \frac{1}{L_{gb}}, & ij\ is\ the\ global\ optimum \\ & and\ L_{gb} is\ the\ shortest\ path \\ 0, & other \end{cases} \end{cases} \quad (1)$$

(2)Local update rules

At the same time, the improved algorithm also added a local update rule that adjusts the information amount on the path during each iteration update, as shown in formula 2. When ants move from node I and node J, pheromone on the path will decrease.

$$\tau_{ij}(0) = \tau_{max}$$
$$\tau_{ij}(t+n) = (1-\rho)g\tau_{ij}(t) + \Delta\tau_{ij}^{min} \quad (2)$$
$$\Delta\tau_{ij}^{min} = \frac{Q}{L},\ L = \min(L_k), k = 1,2,\dots,m$$

(3)State transition rules

The transfer probability of ants is also improved.$p_{ij}^{k}(t)$As shown in formula 3, the transition rule becomes the pseudo-random ratio rule, where it refers to the constant between 0 and 1, and Q is the random number during the period.q_0The next transfer node is determined by comparing the size of the two.That is, if the next node visited by the ant is the node with the maximum value in the equation, otherwise, the transfer probability is still used to judge the next node visited.$q \le q_0$

$$\begin{cases} argmax_{s\in J_k(i)}\{[\tau_{is}(t)]^{\alpha}\cdot[\eta_{is}(t)]^{\beta}\}, & q \le q_0 \\ p_{ij}^{k}(t) = \begin{cases} \frac{[\tau_{ij}(t)]^{\alpha}\cdot[\eta_{ij}(t)]^{\beta}}{\sum_{s\in J_k(i)}[\tau_{is}(t)]^{\alpha}\cdot[\eta_{is}(t)]^{\beta}}, & q > q_0 \\ 0, & other \end{cases} \end{cases} \quad (3)$$

The improved ant colony algorithm initializes the population and each parameter before the algorithm operation, and sets the iteration times.Then the ants were randomly distributed to each node and tabu table were set up.The transfer probability of an ant was randomly calculated according to the transfer probability formula in the improved algorithm, and the next node to be visited was selected. The path distance traveled by each ant in the batch path searching process was recorded, and the shortest path distance and average distance were calculated and recorded. The pheromone increment and pheromone quantity on each path were calculated after the search for the optimal path of this batch of ants was completed, the pheromone on the optimal path was updated, the number of iterations was increased by 1, and the tabu table was cleared.Finally, judge whether the set number of iterations is completed or whether the ant colony has stopped searching. If the algorithm is completed, the optimal path can be obtained.If not, let the colony of the last batch of ants disappear and conduct the path query of the next batch of ants.

D The simulation implementation

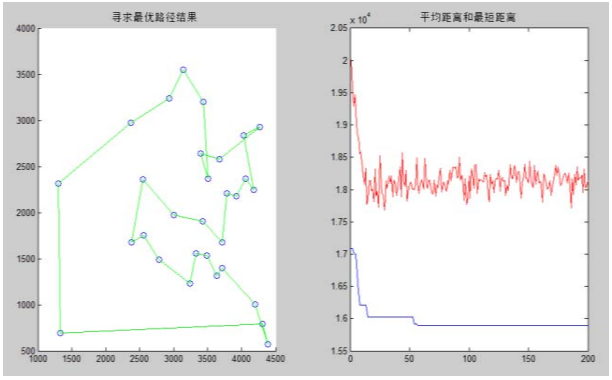


Figure 2 Improved result of ant colony algorithm(1)

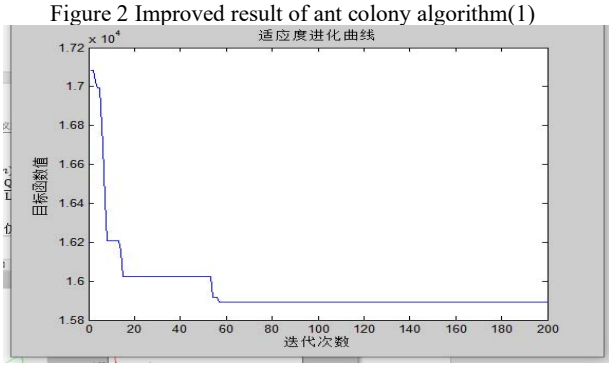


Figure 3 Improved result of ant colony algorithm(2)

As shown in Figure 2, the result of seeking optimal path on the left shows the optimal path result obtained by the improved ant colony algorithm on the map of 4000*4500, while the waveform chart of average distance and minimum distance on the right shows the changes of average distance and minimum distance in the whole algorithm as the number of iterations increases.

As shown in Figure 3, the fitness evolution curve shows that the objective function value of the ant colony algorithm after each iteration, that is, the length of the optimal path changes with the number of iterations, gradually tends to be stable, indicating the accuracy of the improved algorithm for optimal path search.

- *The results are analyzed*

According to the running time and quality of the algorithm, the initial ant number was set as 30, the number of iterations was set as 200, and the pheromone volatilization coefficient was set as 0.5.According to the results in Figure 2 and Figure 3, for large-scale problems, precise methods cannot find feasible solutions within acceptable time, while classical ant colony algorithm and improved ant colony algorithm can find feasible solutions.Acceptable time.Therefore, the improved algorithm provides acceptable solutions to a wider range of problems than the precise method.This shows that the improved algorithm has better optimization ability.The improved algorithm is superior to the original ant colony algorithm, thus reducing the possibility of falling into local optimum.

REFERENCE

[1] Pingping Xia, et al. A Multi-AGV Optimal Scheduling Algorithm Based on Particle Swarm Optimization [A] International Conference on Singapore: Springer, 2020:527-538. (in Chinese)

[2] Drobouchevitch, Inna G., et al. Heuristics for the two-stage job shop scheduling problem with a bottleneck machine[J]. European journal of operational research, 2000, 123(2): 229-240.

[3] Behnam Rahimikelarijani, et al. A mathematical model for multiple-load AGVs in Tandem layout[J]. Journal of Optimization in Industrial Engineering, 2020, 13(1): 67-80.

[4] Tianjian Chen, et al. On the shortest and conflict-free path planning of multi-AGV system based on dijkstra algorithm and the dynamic time-window method[A]. Advanced Materials Research[C]. Switzerland: Trans Tech Publications Ltd, 2013: 267-271.

[5] Nenad Smolic-Rocak, et al. Time windows based dynamic routing in multi-AGV systems[J]. IEEE Transactions on Automation Science and Engineering, 2009, 7(1): 151-155.

[6] Javad Behnamian, et al. Minimizing makespan on a three-machine flowshop batch scheduling problem with transportation using genetic algorithm[J]. Applied Soft Computing, 2012, 12(2): 768-777.

[7] Lacomme P, et al. Resolution Conjointe de problemes d 'Ordonnancement et de la Gestion du Chariot filoguide: Couplage Branch and Bound-Simulation for Evenements Discrets [J]. Congres ROADEF, 2000: 128-129.

[8] Ghada ElKhayat, et al. Integrated production and material handling scheduling using mathematical programming and constraint programming[J]. European Journal of Operational Research, 2006, 175(3): 1818-1832.

[9] Chaudhry I A, et al. Simultaneous scheduling of machines and automated guided vehicles in flexible manufacturing systems using genetic algorithms[J]. Journal of Central South University, 2011, 18(5): 1473-1486.

[10] Nageswara Rao Medikondu, et al. Integration of strategic tactical and operational level planning of scheduling in FMS by metaheuristic algorithm[J]. International Journal of Advanced Engineering Research and Studies, 1(2), 2012.

[11] Ying Chinho, et al. Zone design and control for vehicle collision prevention and load balancing in a zone control AGV system[J]. Computers & Industrial Engineering, 2009, 56(1): 417-432.

[12] Lyle Parungao, et al. Dijkstra algorithm based intelligent path planning with topological map and wireless communication[J]. ARPN Journal of Engineering and Applied Sciences, 2018, 13(8): 2753-2763.

Contourlet Transform Based Reduced-Reference Image Quality Assessment

Fengping Xu
College of information technology
Nantong Normal College
Nantong,China
yctcxfp@163.com

Abstract—**In this paper, we propose a reduced-reference image quality assessment method based on Contourlet transform, due to its properties of multi-resolution, multi-scale, multi-directional and anisotropy. It can capture the image edge profile effectively by using a small number of coefficients. First, we perform three-scale and four-level Contourlet decomposition for the reference and test images. Then we calculate energy eigenvectors on each scale and subsequently obtain the angel of energy eigenvector between the reference and test images. Finally, a weighted summation is used to predict the image quality score. Various experiments based on the LIVE2 database verify the performance of the proposed algorithm, which has high consistency with human visual system.**

Keywords: Reduced-reference Image Quality Assessment; Contourlet Transform; Energy information

I. INTRODUCTION

Distortions and artifacts are inevitably introduced in to digital image during image acquisition, compression, transmission, processing and reproduction. Therefore, it is crucial to access the quality of digital images. There are two kinds of image quality assessment (IQA) methods: subjective and objective image quality assessment methods. As accurate it is, the former method is both time-consuming and energy-consuming, and it cannot be embedded into the system to be assessed automatically, so the latter method is comparatively more viable. Objective image quality assessment method can fall into three categories: Full-Reference (FR), No-Reference (NR) and Reduced-Reference (RR). Up until now of the three methods FR is the most developed, RR less developed, and NR the least for it is still in the primary stage and has not formed a complete effective system. Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) are the most frequently used FR image quality assessment methods. The advantages of MSE/PSNR include: (1) the computational complexity is low; (2) the nice mathematical convexity guarantees a closed-form optimization solution. However, many researchers find that the results they get by means of the method do not correspond with those of the subjective assessments made by people. In [1], it explains in detail why MSE/PSNR is not a desirable predicative method.

Later many scholars have improved image quality assessment methods, and put forward a lot of excellent visual perception methods, for example, the structural similarity index method (SSIM) [2] and wavelet domain reduced-reference image quality assessment algorithm based on natural statistical model [3]. Zhang et al. proposed a feature similarity index (FSIM) [4] that applies phase congruency (PC) and gradient magnitude (GM) in color image quality assessment. In practical RR method is better than FR method because RR IQA methods only need a part of the reference image information and it is more likely to be embedded into real-time application system.

Contourlet transform possesses such properties as multi-resolution, local positioning, multi-directional, neighbor boundary sampling and anisotropy, its basis function distribute in multi-scale, multi-directional, with a small amount of coefficients it can effectively capture the image edge profile, which serves as the major feature of image quality, so it corresponds with human visual system, fits in for image quality assessment, whereby this paper uses Contourlet transform to assess image quality.

The rest of paper is organized as follows: Section 2 addresses the Contourlet Transform theory. Section 3 addresses image quality assessment index. Section 4

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00027

addresses the proposed algorithm. Section 5 addresses the experimental results. Section 6 addresses conclusion.

II. CONTOURLET TRANSFORM

Contourlet transform proposed by Do and Vetterli [5], also known as pyramidal directional filter banks (PDFB), is an efficient representation for image geometry structure. It uses Laplacian Pyramid(LP) decomposition to produce multi-resolution images, and then adopt Directional Filter Bank (DFB) to do direction decomposition on different resolution image, as shown in Fig.1 .In Fig.1,first the circle was decomposed to two scales by LP, as shown in up and down of Fig.1,then the first scale(up) was decomposed to two directions and the second scale(down) was decomposed to four directions by DFB, Contourlet transform was finished eventually.

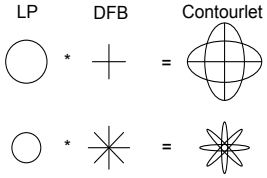


Figure 1. Contourlet Transform Diagram

A. Laplacian Pyramid (LP) Decomposition

Laplacian pyramid decomposition can be used for multi-resolution image analysis. Fig.2 shows the Laplacian pyramid decomposition process, which can be mathematically expressed by:

$$c=Hx,\ p=Gc,\ d=x-p=x-GHx=(1-GH)x \quad (1)$$

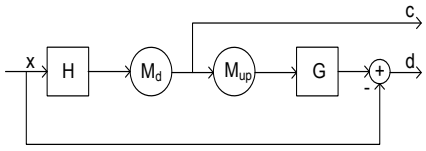


Figure 2. Laplacian Pyramid Decomposition

In Fig.2, LP decomposition uses filter H and down sampling matrix Md to get low pass sub-band (low frequency information) signal c (c = Hx). Then let the low pass sub-band signal c pass the up-sampling matrix Map and synthetic filter G to get forecast signal p (p = Gc). Finally, we can get the difference signal d (d = x - p) (high frequency information) between the original image signal and forecast signal. Apply the process stated above to next layer, we can get multi-resolution images.

B. Directional Filter Bank

After LP decomposition, we can get the high frequency information of different direction by using the Direction Filter Bank (DFB), DFB includes two modules: double-channels Quincunx filter bank and shearing operation. The Quincunx filter has two levels, where the first level decomposition outputs vertical directional sub-band and horizontal directional sub-band. The second level decomposes y0 into y00 and y01, and decomposes y1 into y10 and y11. From the third level, we first do shear operation, and then use Quincunx filter. The third level DFB decomposition of y00 and y01 can be divided into two types according to the upper channel 0 and the down channel 1. The upper channel uses Type-1 decomposition, the down channel uses the Type-2 decomposition. The y10 and y11 also have two DFB decomposition forms, which are similar to the y00 and y01.

The relationship between the number of image decomposition levels and the number directional sub-bands is M=N2, where N is the number of decomposition levels on each scale, M is the No. of obtained sub-bands on each scale.

III. IMAGE QUALITY EVALUATION INDEX

Take the famous zoneplate picture, as an example (Fig.3). We do 3-scale Contourlet decomposition to it and subsequently 4-level directional decomposition on each scale. Thus, we can get 16 different directional sub-bands on each scale, as shown in Fig.4. We define the sum of squared coefficients to be the sub band energy: $Eng = \sum_{i=1}^{M}\sum_{j=1}^{N} C(i,j)^2$, where C (i, j) is the sub band coefficient of row i column j, M and N is the resolution of the scale images respectively. Then, we can get the energy feature vector of each scale sub band $Eng_{vector}(k)$:

$$Eng_v(k) = ln\big(Eng(k)\big) \qquad k=1,2,\cdots,16 \quad (2)$$

where k is the k-th directional sub band.

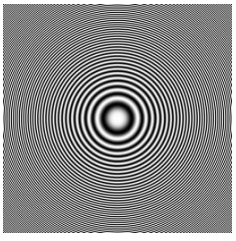


Figure 3. 6zoneplate

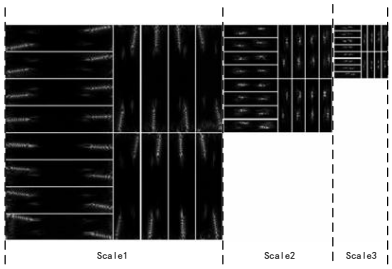


Figure 4. Contourlet Decomposition for zoneplate

The angle between two images' energy vectors can be used to measure the similarity of them:

$$\theta = arccos \frac{<Eng_vX, Eng_vY>}{\sqrt{<Eng_vX, Eng_vX>} * \sqrt{<Eng_vY, Eng_vY>}} \quad (3)$$

Where <x, y> is the dot product. EngvectorX and EngvectorY are 16 dimensional energy feature vectors of reference image and distortion image respectively.

If the angle between the distortion image energy feature vector and the reference image energy feature vector is smaller, the distortion image quality is better. According to this principle, we define the image quality evaluation by using the angle between the energy feature vectors of reference image and distortion image on each scale:

$$QENG = w_1 * \theta_1 + w_2 * \theta_2 + w_3 * \theta_3 \quad (4)$$

Where $\vec{w}$ is the weight of each scale, and $\vec{\theta}$ is the energy feature vector angles between the reference image and distortion image on scale1, scale2 and scale3, they are calculated by formula (3), and w_1 , w_2 and w_3 respectively is weight on each scale.

IV. Reduced-Reference Image Quality Evaluation Algorithm Based on Contourlet Transform

The proposed Contourlet transform based Reduced-Reference image quality evaluation algorithm can be summarized follows:

Step 1: Apply 3 scale 4 level Contourlet decomposition for distortion image and the corresponding reference image, and get 16 different directional sub band on each scale.

Step 2: Calculate energy feature vectors $Eng_v(k)$ on each scale for the distortion image and reference image respectively.

Step 3: Calculate the energy feature vector angles θ_1,θ_2 and $\theta_3$between the distortion image and reference image on each scale by using (3) respectively.

Step 4: Calculate the image quality value QENG by using $\vec{w} = [0.1, 0.2, 0.7]$.

V. Evaluations

A. The Experiment Database and Evaluation Index

To validate the effectiveness of proposed algorithm, we test it on the LIVE Image Quality

Assessment Database Release2(LIVE2) from the University of Texas [6]. The LIVE2 database has 5 kinds of distortion type images, including Fast Fading image, Blur image, JPEG compression image, JP2K compression image and Noise image, as shown in table 1. It provides "subjective Difference score", Difference Mean Opinion Scores (DMOS), for all distortion images, which describes the difference between subjective score (Mean Opinion Scores, MOS) and full score 100 (namely DMOS = 100 - MOS). Thus, the bigger DMOS is, the poorer image quality is.

TABLE I. LIVE2 Image Database

Distortion type	Distortion mages	Source Images
FastFading	145	29
Blur	145	29
JPEG	175	29
JP2K	169	29
Noise	145	29

We choose two common objective parameters as our evaluation index: Correlation Coefficient (CC) under the condition of nonlinear regression and Spearman Rank Correlation Coefficient (SROCC). CC and SROCC are the linear correlations between objective quality and subjective quality under the nonlinear regression conditions. The higher the value is, the better the correlation of objective evaluation method and subjective quality is. SROCC and CC are between 0 and 1. The closer to 1 their value is, the better the performance is. CC value is calculated by nonlinear regression function of five parameters $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$in literature [6-7]:

$$Quality(x) = \beta_1 logistic(\beta_2, (x - \beta_3)) + \beta_4 x + \beta_5 \quad (5)$$

$$\text{logistic}(\tau, x) = \frac{1}{2} - \frac{1}{1+\exp(\tau x)} \quad (6)$$

B. *Analysis of Experimental Results*

Our proposed algorithm was compared to several state-of-the-art RR-IQA and FR-IQA metrics, including WNRS [3], SSIM [2], SUMMER [7], CEQI [8], and VCGS [9], whose original source code are available online. The experimental results have two parts: CC and SROCC value comparison of different algorithms and DMOS value of different distortion types by using the proposed algorithm, as shown in table 2.

TABLE II. COMPARISON OF DIFFERENT ALGORITHMS ON LIVE2 DATABASE

Measure	Model	FastFading	Blur	JPEG	JP2K	Noise
CC	WNRS	0.9263	0.8883	0.8760	0.9245	0.8898
	SSIM	0.9492	0.8741	0.9297	0.9368	0.9793
	SUMMER	0.9663	0.9683	0.9560	0.9545	0.9698
	CEQI	0.9572	0.9702	0.9623	0.9345	0.9645
	VCGS	0.9567	0.9781	0.9687	0.9676	0.9723
	proposed（RR）	**0.9706**	**0.9612**	**0.9789**	**0.9689**	**0.9436**
SROCC	WNRS（RR）	0.9229	0.9147	0.8508	0.9205	0.8702
	SSIM（FR）	0.9411	0.8943	0.9107	0.9317	0.9629
	SUMMER	0.9512	0.9678	0.9534	0.9598	0.9665
	CEQI	0.9634	0.9711	0.9656	0.9456	0.9678
	VCGS [69]	0.9556	0.9638	0.9707	0.9621	0.9764
	proposed（RR）	**0.9637**	**0.9710**	**0.9658**	**0.9479**	**0.9660**

In Table 2, CC and SROCC are the correlation coefficient used to measure the performance of image quality evaluation algorithms. The value is between 0 and 1. If it is close to 1, the image quality evaluation index is highly correlated with DMOS, and the performance of image quality evaluation is better. It can be seen from Table 2 that the proposed algorithm has best evaluation index on all 5 kinds of distortion type images. The index value is stable and suitable for all kind of distortion. For full reference algorithms, the performance of the proposed algorithm is slightly less than the rothers for noise distortion type. But it performs better in blur distortion type, JPEG distortion type and JP2K distortion type. Also, the proposed algorithm has high consistency with the human visual system.

In addition, the proposed algorithm is a kind of reduced reference image quality evaluation method. Compared with full reference image quality assessment algorithm, we only need a part of the information of images. Furthermore, the proposed algorithm has advantages including transport easily, easy to be embedded into real-time application system and suit for practical application.

VI. CONCLUSION

The main function of Contourlet transform is multi-scale and multi-directional distribution with the characteristics of multi-resolution, locality and directionality. Therefore, a small number of coefficients can effectively capture the edge contour of the image, conforming to the characteristics of human vision and suitable for image quality evaluation. The image is decomposed by 3 scale and 4 level Contourlet, the included angle of energy eigenvectors of each scale is calculated, and the weighted sum is carried out to obtain the image quality evaluation standard QENG in this paper. The smaller QEBG value is, the better image quality is. The experiment results show that the algorithm has good performance, and is more superior performance than other literature algorithm. No reference image quality evaluation method will be the future research direction, which extracts image quality characteristics based on Contourlet transform.

REFERENCE

[1] Z. Wang and A. C. Bovik, “Mean squared error: Love it or leave it? - A new look at signal fidelity measures,” IEEE Signal Processing Magazine, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[2] Z Wang, A. C. Bovik and H. R. Sheikh, “Image Quality Assessment: From Error Visibility to Structural Similarity,” IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 13, NO. 4, APRIL 2004.

[3] Z. Wang, E. P. Simoncelli, “Reduced-Reference Image Quality Assessment Using A Wavelet-Domain Natural Image Statistic Model,” in Proc. of SPIE Human Vision and Electronic Imaging, vol. 5666, Mar. 2005, pp. 149-159.

[4] L. Zhang, X. Mou, “FSIM: A feature similarity index for image quality assessment,” IEEE transactions on Image Processing, 2011, 20(8): 2378-2386.

[5] M. N. Do, M. Vetterli, “The Contourlet Transform: An Efficient Directional Multiresolution Image Representation,” IEEE TRANSACTIONS ON IMAGE PROCESSING, vol.14, no.12, pp.2091-2106, Dec. 2005.

[6] H. R. Sheikh, Z. Wang, L. Cormack, “Live image quality assessment database release2,” 2004.http://live.ece.utexas.edu/research/quality.

[7] D. Temel, G. AlRegib, “Perceptual image quality assessment through spectral analysis of error representations,” Signal Process, Image Commun. 2019, 70, 37–46.

[8] M. Layek, A. Uddin, T. P. Le, T. Chung, E. N. Huh, “Center-emphasized visual saliency and a contrast-based full reference image quality index,” Symmetry 2019, 11, 296.

[9] C. Shi, Lin, Y. Full, “Reference Image Quality Assessment Based on Visual Salience with Color Appearance and Gradient Similarity,” IEEE Access 2020.

Emotional Conversational Analytics – An experimental case study

Dr. Jay Kiruthika
Kingston University
j.kiruthika@kingston.ac.uk

***Abstract*— Users are shifting to voice over message-based communications rather than typing. There exists a gap in evaluating emotional aspects of the phrases uttered. There are software analytics for sentimental analysis and other ways of measuring emotional aspects of the spoken phrases. But there is a lack of a system that evaluates secondary and primary emotions that can be used in various areas to detect human behavior. In order to facilitate this, there is a need to decipher the primary and secondary emotions of the user. This project addresses a novel idea of creating an emotional phrase map that identifies secondary and primary emotions, in addition to using machine language to train the system to identify the phrases and classify the emotions via a case study and the results analyzed.**

Keywords- Emotional Conversational analytics; Emotional Conversations; Emotional NLP;

I. INTRODUCTION

The global market for voice search devices grows every year as the devices are commercialized for budgeted households. Companies like Apple, Google, Xiaomi, Amazon and other manufacturers have flooded the market for this purpose. These devices have various functions and features that facilitate a user to voice search, find information, retrieve everyday planner, do video calling, communicate with each other etc. Users like to use voice communication as it is faster and need not be tied to a physical keyboard/device. The voice commands can be uttered in a physical area of coverage. Often sentiment scoring is used to determine primary emotions of the uttered words by the users. But secondary emotions are ignored. Using Robert Plutchik's[1] emotional wheel, primary and secondary emotions can be evaluated to learn more about emotional aspects of the voice chat or conversation. As future growth of these devices is expected to grow exponentially there is a need to elucidate emotions of the phrases uttered. This also can be used for monitoring and other purposes.

This project tends to use voice over chat conversations to provide intelligent conversational analytics solutions intended for various uses. For example, it can be used to identifying vulnerable users seeking help which otherwise is not identified by other means. It can also be used in detecting cyber threat/violence and abusive behavior. A smart notification system can be built on the resulting analytics that can be used to monitor the user's intent. This requires creating a new type of phrase map(s), named emotional phrase map(s) in this project that can help identify the primary and secondary emotions of the phrase uttered (Fig 1.1).

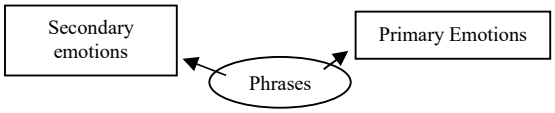


Fig 1.1 Learning Journey.

The NLP and NLU are currently used to decipher phrases, focus on understanding the language structure in addition to determine user's intent. Google, Amazon uses training data to achieve a higher success rate in determining user intent. Plutchik[1] classifies primary & related secondary/ and the dyads of emotions that sit between two. This can be integrated to current NLP for deciphering human emotions.

Primary	Secondary	Sub sect of Secondary	Dyads of the subsect
Rage	Anger	Annoyance	
Vigilance	Anticipation	Interest	Aggressiveness
Ecstasy	Joy	Serenity	Optimism
Admiration	Trust	Acceptance	Love
Terror	Fear	Apprehension	Submission
Amazement	Surprise	Distraction	Awe
Grief	Sadness	Pensiveness	Disapproval
Loathing	Disgust	Boredom	Remorse
Rage	Anger	Annoyance	Contempt

Fig 1.2 Emotional classification with dyads

The current focus of an NLP[2] is to :

- Understand language semantics
- Understand User Intent
- Understand the phrases with variable synonyms etc via phrase maps.

II. EMOTIONAL PHRASE MAPS

The phrase maps are used to recognize and register the words that were being uttered by a user. This helps to understand the communication between a human and a

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00028

device to provide meaningful feedback to the user. These phrase maps are predominantly used in creating voice bots and in CUI(Conversational User Interface) designs. A regular phrase map[4][6] contains insightful text analysis with machine language that extracts, analyzes and stores text. This can be used to train & customize applications and to classify, extract and detect sentiment with minimum effort using NLU[7] resulting in sentiment analysis, entity analysis, infographics, content classification etc.

Whilst creating an emotional conversational analytics platform it has been found that these phrase maps are not useful as it does not include the emotions of the user. A new emotional phrase map can be created such as one listed (Fig 2.1)

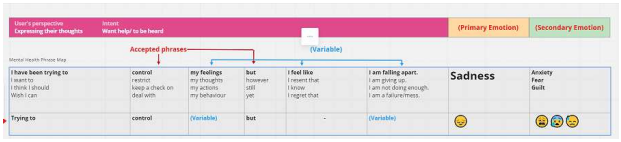


Fig 2.1 Emotional Phrase Map for mental health

This is a new design artifact that includes the primary and secondary emotions of the user. The last two columns have been added to create this phrase map. By using these labels it can automatically create a sub token that can be mapped to NLU/NLP system to create the emotional map. In the above diagram this phrase was uttered by a user with mental health issues and is a simulated text. This can be used in various scenarios to produce infographics for an organization, social media app etc. By deciphering the user's emotions their behavior can be analyzed and actions if needed, can be taken before it escalates. It can also be adopted to detect cyber threats targeting vulnerable individuals, mental health issues, organizational abuse, mental well being in assisted living care, domestic violence, threats etc.

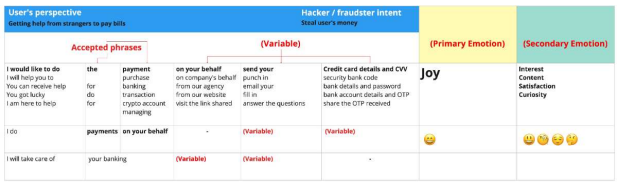


Fig 2.2 Emotional Phrase map for cyber threat towards vulnerable user

There are anomalies in emotional phrase map as it maps the user's emotions. In Fig 2.2 the user shows joy as there is a person volunteering to help with his payments. This is again a simulated text showing how a vulnerable user will react. This intent is again indeed harmful to the said user. This needs further research and classification of emotions in such situations.

The various dimensions of gathered data based on emotions can be used to provide crucial analytics to improve customer experience, emotional wellbeing, monitoring tool for assisted living etc. paving way to new products and solutions.

III. CHALLENGES

Emotional Conversational analytics: Human emotions are varied as per the situation or the personality of the individual. The sentiment analysis[13] and scoring gives a good idea about negative, positive or neutral feelings based on the keywords uttered. They are rated as per the product the users react to and utter. Deciphering primary emotions has improved tremendously with the existing NLP as it is able to classify them. But the secondary emotions need to be analyzed extensively as the primary emotions related to it might not be necessarily the same. In a conversation, a sarcasm can be deciphered differently. For example, instead of I love you the user might utter I hate you, if it is taken in a perceived sense[9] then the emotions related to it will differ.

User Intent: User intent in emotions is one of the core challenges any emotional conversational analytics platform will face. The emotions related to certain phrases might differ based on culture[11], personality of the user, native speakers, slangs[15]. This is true for any human conversation. The complexity of the conversation increases whilst spoken out of context in occasions. The result may vary depending on the audience. But deciphering such emotions based on the user intent is a challenge.

Algorithms: Neural network[14], transfer learning[12] are few of the algorithm that is suitable to decipher emotions and language processing. There is a need to increase the accuracy based on hypothetical classification of emotions. This may vary, as there is a need to set the classifications to be a common base before attempting to further sub classify secondary emotions. The local variations of user intent and emotional response needs to be evaluated in a meaningful manner if the system is to be used efficiently.

Slangs: How far can the user's private slangs hold against normal human conversations? Especially in restricted communities' humans used to substitute phrases with words that is known inside their community and is understandable. But if a system is to decipher these slangs a complex system is needed to decipher the emotional aspects of that conversation. The library then needs to add these slangs as acceptable sub phrase(s).

IV. CASE STUDY- EXAMPLE SCENARIO

For research purpose, multiple scenarios are considered to decipher user's emotions. Cyber threat, Mental wellbeing, Organizational abuse, domestic abuse was focused on this case study and corresponding emotional phrase maps were

designed. The training dataset of the phrases used in this project is from Sentiment140. An online tool was built to label existing sentences on emotions.

The initial step is to label the phrases, this was done manually and in the future version once the system is trained it can be automated. The text is cleaned of emojis/slangs then phrases were labelled(Fig 4.1)

Fig 4.1 Table of primary and secondary emotions

Transfer learning algorithm was considered in this case study for layering various sentiments across the phrases. It deciphers primary and secondary emotions on the training data. Languages form word roots with structured verbs, adjectives etc. The stemming finds the word extension and removes it. Lemmatization knows the word as such and brings it back to its root via a lexicon. This reduction technique is vital whilst incorporating emotion in connection with a word sequence; certain roots and extensions carry the same emotion.

Therefore, it allows us to reduce the number of variations to its strict minimum to recognize an emotion. It also helps to reinforce the impact of a word for emotion recognition and to standardize a text in a consistent way. Example: love, loving, lover all corresponds to love. NLTK library which offers both stemming tools and lemmatization tools was chosen[3]. For the lemmatization WordNet word bag is used to assign the words to treat tags like "Adverbs" or "Pronouns" so that they receive the corresponding root. As a result, lemmatization[5] requires more preparation to be implemented than stemming which simply truncates the word and singularizes it. The 2 approaches offer quite similar results. Although lemmatization has slightly higher precision and is also increasingly preferred in NLP projects.

In this case study a generic model with 28 different labels as input was used to classify the text. This method took time to train and offered very poor results. This led to creation of several models and each one specializing on one type of emotion. Considering this, the final product is an algorithm constituting different types of layers in the model with the following architecture:

• 1st layer: “Colours” such as Neutral, Yellow, and so on (Fig 4.1)

• 2nd layer: “Sub Colours” which deals with the emotion inside of a colour for example:

If the 1st layer found a higher probability of “Yellow” then the 2nd layer will be between the 3 emotions inside such as Serenity if its probability is higher than the other 2. This layering architecture offers a possibility to tune finely each Colour’s model. In addition, Neutral colour is the most represented in our dataset and its model tended to not achieve great results. To counter this issue, few strategies were explored to improve its training quality:

1. Consider that secondary emotions are “in between” emotions, in the case 2 “Colours” has been predicted and both are neighbours (thus having a common secondary emotion), it is concluded that the secondary emotion is the right prediction. In case of 3 “Colours”, higher probability is chosen.

2. Adding secondary emotion classifier models in the 1st layer, and simply choosing the highest probability for the prediction.

The “Colours” layer contains 9 binary classifier models or 17 depending on the strategy used, which prevents emotion from overlapping. The “Sub Colours” layer contains 8 classifier models defining probabilities for 3 emotions per model. The resultant accuracy and loss are analyzed and listed(Fig 4.2)

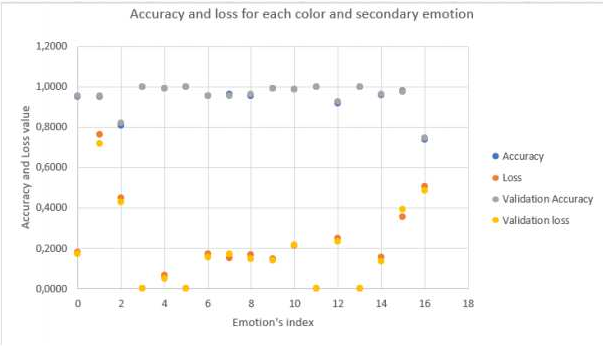


Fig 4.2 Accuracies and losses for each primary and secondary emotion(s)

IV Conclusions and Future Work

The growth of voice devices is increasing and there is a need to include human emotions into the system to address various issue to make it natural and humanlike behaviour[8]. It can also be used in various discipline to produce

intelligent analytics. For example, an organizational violence happening in a region or an area of a MNC can be identified via infographics once the emotional intent of the user has been deciphered. This can be used to create smart notifications that can be used for prevention and action can be taken as per the nature of the issue.

The mental wellbeing of a patient or residents in care homes can also be determined and actions can be taken accordingly. Organizations are implementing voice bots/ robots that greet and cater to the user needs as per the user requests. The robots read facial expressions to respond in a realistic manner. As the voice over devices usage is increasing year on year, there is a need for such a system to humanize the devices.

The future of this application can be adopted to reality technologies in everyday activities and implemented to support independent living including vulnerable elderly by auto monitoring their emotions and address these needs before they escalate. Further research is needed on analysing tone to detect aggressive behaviour and other aspects of human emotions that can be incorporated into NLP and NLU. This system can be adopted to personalize voice chatbots/ virtual assistants to determine emotional needs of the user and cater accordingly.

As more and more of the users are migrating to voice over devices, it is imperative to add emotional recognition [10] into the NLP and NLU for better understanding and simulating human behaviour in systems. Especially this is prevalent with vulnerable and elderly who otherwise refrain from using technology.

These voice devices pave the way for users to access information easily without learning to use the interface that needs typing. There is a need to train the system to improve the accuracy as this research is in the early stages and requires a larger data set to determine anomalies. The accuracy rate for the case study is 24% and needs further research in providing cutting edge analytics based on the data gathered. It can also promote “humanlike” behaviour in the systems.

The wide spread of such technology can change the way/behaviour of the consumers and businesses opening to newer possibilities of deep personalization of their services and challenges to contend with.

ACKNOWLEDEMENT

I sincerely thank Prof. J.C. Nobel, Kingston University, Gaëtan Barret, Himanshu Bhasker, Shubhra Singh, Harshil Shah, Mokshi Charunika Dangampolage for their assistance.

REFERENCES

[1] Plutchik R., Kellerman H. 2013 Theories of Emotion. Volume 1 of Emotion, theory, research, and experience, Academic Press,1483270017, 9781483270012

[2] Sabharwal N., Agrawal A. 2021 Introduction to Natural Language Processing. In: Hands-on Question Answering Systems with BERT. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-6664-9_1.

[3] Millstein F. 2020. Natural Language Processing With Python: Natural Language Processing Using NLTK

[4] Withanage, P., Liyanage, T., Deeyakaduwe, N., Dias, E., & Thelijjagoda, S. 2018. RoadNavigation System Using Automatic Speech Recognition (ASR) And Natural Language Processing (NLP). 2. I. (R10-HTC) (Dü.). içinde Malambe: IEEE. doi:10.1109/R10-HTC.2018.8629859

[5] Po-Hao Chen, 2020. Essential Elements of Natural Language Processing: What the Radiologist Should Know, Academic Radiology, Volume 27, Issue 1,Pp 6-12,ISSN 1076-6332,https://doi.org/10.1016/j.acra.2019.08.010.

[6] Thejaswee M., Srilakshmi V., Anuradha K., Karuna G. 2021. Performance Analysis of Machine Learning Algorithms for Text Classification. In: Luhach A.K., Jat D.S., Bin Ghazali K.H., Gao XZ., Lingras P. (eds) Advanced Informatics for Computing Research. ICAICR 2020. Communications in Computer and Information Science, vol 1393. Springer, Singapore. https://doi.org/10.1007/978-981-16-3660-8_39

[7] Emily M. Bender, Alexander K. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.

[8] Harari, Y. Reboot for the AI revolution. Nature 550, 324–327 (2017). https://doi.org/10.1038/550324a

[9] Xiaojuan Ma, Emily Yang, Pascale Fung. 2019. Exploring Perceived Emotional Intelligence of Personality-Driven Virtual Agents in Handling User Challenges. WWW '19: The World Wide Web Conference. https://doi.org/10.1145/3308558.3313400.

[10] Valentina Ferretti, Francesco Papaleo. 2018. Understanding others: Emotion recognition in humans and other animals, Wiley online library. https://doi.org/10.1111/gbb.12544

[11] Arvid Kappas, Nicole C. Krämer. 2011. Face-to-Face Communication over the Internet: Emotions in a Web of Culture, Language, and Technology, Cambridge University Press

[12] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015, pp. 1225-1237, doi: 10.1109/APSIPA.2015.7415532

[13] R. Liu, Y. Shi, C. Ji and M. Jia, "A Survey of Sentiment Analysis Based on Transfer Learning," in IEEE Access, vol. 7, pp. 85401-85412, 2019, doi: 10.1109/ACCESS.2019.2925059.

[14] Mohamed Morchid. 2019. Neural Networks for Natural Language Processing. Artificial Intelligence [cs.AI]. Avignon Université, fftel-02398814v2f.

[15] Zhewei Sun, Richard Zemel, Yang Xu; 2021. A Computational Framework for Slang Generation. Transactions of the Association for Computational Linguistics; Vol.9 pp462–478. doi: https://doi.org/10.1162/tacl_a_00378.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Application of face recognition in port unrestricted scene

Song Qi
School of Computer Science and Technology,
Wuhan University of Technology
Wuhan, China
e-mail: 1192913039@qq.com

HanBing Yao
Wuhan University of Technology
Chongqing Research Institute
School of Computer Science and Technology,
Wuhan University of Technology
Wuhan, China
e-mail: 22876681@qq.com

WuWenBo
Imformation Technology Department,
Jiangjin Port Branch
Chongqing Gangjiu Co., LTD.
ChongQi China
e-mail:441479417@qq.com

***Abstract*—Traditional face recognition technology needs to collect data in a specified environment with stable facial features.However,due to the large scope and wide area of port operation,traditional identification methods have been unable to meet the requirements of port unrestricted scenarios. In this paper,the algorithms based on RetinaFace face detection and ArcFace face recognition are applied to port staff attendance in unrestricted scenarios.In order to obtain a more stable and accurate face representation method,this paper studies face detection, face image pretreatment,face recognition,loss function and other aspects,and verifies the feasibility of this model in port unrestricted scene attendance through the experimental results on PubFig dataset and the analysis of real-time face recognition effect.**

***Index Terms*—data sets;Face recognition;ArcFace;Loss functions;**

I. Introduction

Nowadays, with the trend of world economic globalization,the operation of port enterprises is also developing towards a more professional and information mode.A good attendance system not only enables enterprise leaders to better manage employees,but also enables employees to better restrain themselves,put more time and energy into work, and bring better development for the enterprise.Face recognition is a research hotspot in computer vision field.The traditional face recognition methods are only suitable for some specific scenes,such as the access board of public facilities.In general, only when the facial features reach a stable state, the traditional face recognition technology can accurately detect,and also require the detection of the face recognition equipment,the detection of active cooperation,such as removing masks,glasses and other objects that block the face recognition.The attendance method based on face recognition eliminates the disadvantages of traditional attendance,greatly facilitates personnel management,and makes attendance more reliable and efficient. Face recognition is usually composed of three key elements:face detection, face pretreatment and face recognition.The system designed in this paper uses face detection algorithm RetinaFace and face recognition part uses an additive angular edge loss (ArcFace) algorithm.However,the combination of Multi-task Cascaded Convolutional Neural Networks(MTCNN)face detection algorithm and Facenet face recognition algorithm had been better in the actual test environment.

II. FACE RECOGNITION

According to the different forms of data, face recognition can be divided into methods based on 2D images and methods based on 3D scanning,the two methods are very different in the development and application to a certain extent.In addition, with the rapid development of deep convolutional neural networks (DCNNs),deep learning-based methods have achieved significant performance improvements in including face recognition.In general, the depth face recognition system includes three basic elements:face detection, face pretreatment and face recognition.First,face detection locates the face region on the input image.Then face preprocessing is carried out to normalize the detected face into a standard layout. Finally,the face recognition algorithm will extract the recognition features from the preprocessed face.These features were used to calculate how similar they were to each other to determine whether faces belonged to the same identity.

A. Face detection

Face detection is the first step of face recognition system.Its main task is to lock the face region in the picture,for example, given an input picture,the task of face detection is to find all the faces in the picture, and give the boundary box coordinates with confidence.Face detection technology has made great progress in recent years,but there are still some difficulties and challenges,such as scale,resolution,occlusion,illumination and so on.Traditional face recognition technology designed manual features are mainly used to distinguish between face and background area.The method can be divided into single-stage and multi-stage methods by stage learning, and the anchored based and non-anchored methods can be divided according to the usage of anchor.However, a face detection method can be both single-phase and anchor-based.For the multistage approach,the multistage detector first generates many candidate boxes and

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00029

then refines them in a coarse-to-fine or coarse-to-refinement strategy.

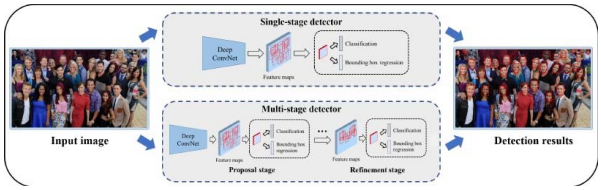


Fig. 1. Schematic diagram of single and multistage face detector

As shown in Figure 1,the single-stage detector directly completes face detection from the whole feature graph,while the multi-stage detector uses multi-stage to generate candidate images and refine candidate images through one or more stages. A classic single-stage detector,SSD, is the mainstream framework for target detection because it runs faster than the faster R-CNN while maintaining considerable accuracy. Many developers are using SSD face detection in the system. Although single-stage method has the advantage of high efficiency, its detection accuracy is lower than multi-stage method. This is partly due to the negative and positive imbalance caused by dense anchors.Due to the long development time and superior performance, most face detectors are based on anchor points. Dense anchor points are usually preset on the feature graph, and then these anchor points are classified and boundary box regressions are performed one or more times.

B. Face preprocessing

In the second stage,the purpose of face preprocessing is to calibrate the detected face to a typical view (mainly to improve the end-to-end performance of face recognition is an important step.Because the structure of the face is regular, and face parts (eyes, nose, mouth, etc.) of the arrangement is constant,so the arrangement of the face of the subsequent face recognition feature calculation has great benefits.In general, face alignment uses spatial transformation techniques to calibrate faces into a standardized layout.In the existing face alignment methods,face landmarks or so-called face keypoints are essential because they serve as a reference for similar transformation or affine transformation.Therefore, face landmark localization is the premise of face alignment.Face landmark localization methods based on deep convolutional neural network can be divided into three categories: coordinate regression based method,heat map regression based method and 3d model fitting method.RetinaFace puts the face image on the ordinate of the landmark based on the return target.The study of coordinate regression mainly adopts L1, L2 or smooth L1 loss function.

C. RetinaFace

RerinaFace is a single-stage target detection algorithm specialized in face detection.Based on multi-stage,Feature Pyramid Networks (FPN) are used to make full use of Feature information.The network structure is shown in Figure 2.RetinaFace uses multilevel feature images to form a feature pyramid,and a large number of anchor boxes with different sizes are designed on the multilevel feature images to achieve excellent detection performance.

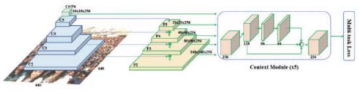


Fig. 2. RetinaFace algorithm network structure

D. Face recognition

After the face preprocessing,in the face recognition stage,the goal is to map the aligned face image to a feature space,where the features of the same identity are close,and the features of different identities are far apart.In practical application,face recognition mainly has two tasks:face verification and face recognition.But whether it is face verification or face recognition,are used to calculate the similarity of face image features,how to learn different face representation is the core goal of face recognition system.[1]Network architecture and training supervision play a key role in face recognition learning.

The supervised training methods of face recognition are divided into supervised mode,semi-supervised mode and unsupervised mode. In the supervision mode,it can be divided into three categories:classification,feature embedding and hybrid method. Deep face recognition learning based on classification method comes from general object classification task.Each class corresponds to an identity that contains multiple faces of the same person.Softmax training loss is the most widely used supervision method in classification tasks, which consists of full connection layer, Softmax and cross entropy loss.

E. ArcFace algorithm

SoftMax is often used in the classification process of various categories in deep learning.It allows the output values of multiple neurons to be mapped between 0 and 1 and their sum is 1.With similar probability,we take the node with the largest output value as the target category for prediction.Softmax cross entropy loss function:it is often used in the classification process of various categories in the past.[3]It allows the output values of multiple neurons to be mapped between 0 and 1 and their sum is 1.With similar probability,we take the node with the maximum output value as the target category for prediction.Softmax cross entropy loss function formula is shown in Formula 1:

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{w_{yi}^T xi+by_i}}{\sum_{j=1}^{n} e^{w_j^T xi+bj}} \tag{1}$$

Xi is the sample feature,Yi is the category to which the sample belongs,Wj is the jth column of the weight matrix,bj is the paranoid amount,N is the batch processing size,and N is the number of categories.The unimproved SoftMax loss function can only classify them by output value,and cannot closely combine the same category through optimization,nor can it

distinguish different categories greatly,which is not suitable for face recognition training.ArcFace moves the additive Angle margin M to the interior of cos,and the loss function formula is shown in Formula 2:

$$L4 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta yi+m))}}{e^{s(\cos(\theta yi+m))}+\sum_{j\neq 1,j\neq yi}^{n}e^{s\cos\theta yi}} \quad (2)$$

Compared with other loss functions,ArcFace can better interpret Angle distance in geometric space,and it can correspond to geodetic lines in hyperplane and maximize decision boundary more intuitively.

III. EXPERIMENTATION

A. Experimental data set

The experimental Dataset PubFig is the Columbia University Open People Dataset.It is a large face dataset that covers 58,797 images of 200 people on the Internet.Unlike most existing face datasets,these images are taken with the subject completely out of control.Therefore,there are great differences in posture,illumination, expression,scene, camera,imaging conditions and parameters in different images. Through data set provides web crawl the data link,which the web page is inaccessible data cleaning,do not use data,not the camera captured data and light discomfort,mask,the image is not clear,such as data, lastly,everyone a few different positions (as shown in figure 3),facial expression, illumination and so on,Four images per person were used as validation sets,which contained 1 200 positive sample pairs and 398 000 negative sample pairs.Before training,face detection and clipping of the dataset with RetinaFace,image data preprocessing clipping size.See Figure 3

Fig. 3. Partial data set

The data set training in this paper is based on Ubantu20.04, Pytorch deep learning framework, and trained on two GeForce GTX 1080Ti GPU. The skeleton structure of the network is ResNet50 in deep convolutional neural Network. Arcface is used for the loss function, and stochastic gradient descent (SGD) is adopted. In order to make the network converge faster, cosine annealing attenuation method is used in the training. The initial learning rate is set to 0.1, and the maximum number of iterations is set to 100. After 100 batch training, the network reaches convergence state.Due to the small data set, the convergence speed of the loss function is fast, and the accuracy of the final retraining set and verification set is high. The final experiment also achieved the expected effect.

As can be seen from Table 1,after 10,000 batch training of the algorithm,ArcFace has the highest accuracy,SphereFace has the lowest accuracy than ArcFace,SoftMax performs poorly compared with other loss functions,CosFace and ArcFace have little difference.It can be seen that the performance of ArcFace loss function has a good performance in PubFig.The convergence of each loss function is good during the model training.Due to the small size of the data set and the high accuracy of training and verification,the loss function converges rapidly and achieves good results.It can be seen that the experiment on the PubFig public data set has achieved the expected goal.

TABLE I
ACCURACY OF DIFFERENT LOSS FUNCTIONS

loss function	Highest accuracy (percent)
CosFace	82.62
SphereFace	81.24
SoftMax	70.92
ArcFace	87.00

IV. PORT STAFF ATTENDANCE FACE RECOGNITION APPLICATION

A. Application Background

The common way of attendance is radio frequency card attendance and fingerprint attendance.These two methods of attendance are also widely used, but both card attendance and fingerprint attendance face the problem of whether the puncher is himself or not.Radio frequency card attendance can be handed to other people to clock in,while the fingerprint attendance mode seems to be unable to clock in,but now fingerprint film attendance technology has been very common,can not avoid the phenomenon of clock in. With the advent of the facial recognition system,the phenomenon of punching in has been solved,and it is not possible to use photos to punch in.As the facial recognition technology becomes more and more mature,the accuracy of facial recognition is gradually improved to achieve fast facial recognition.It has been a trend for companies to use face recognition attendance systems.

The working environment of ports is large,and the working time and location of employees are not uniform,which is quite different from that of ordinary enterprises.As a result,most port enterprises still use the most traditional paper and handwritten attendance method. On the one hand, port employees, this special group,are engaged in port operations, and a large number of port employees have incomplete fingerprint information,resulting in a large number of port enterprises in the use of fingerprint attendance is powerless; On the other hand,most domestic researches on ports focus on port intellectualization and informatization,and less attention is paid to port enterprise staff attendance management.Therefore, the mobile attendance mode of face recognition is also one of the attendance modes relatively suitable for this group.

B. The overall design of port face recognition system

The specific implementation steps of the system are as follows:

1) Login module : In order to ensure the security of employee information, the system has a login module, only authorized administrators can log in operations. This module is connected to the MySQL database. The administrator must enter the correct user name and password to log in successfully. After login successfully, the administrator can view the basic personal information and attendance information of employees in the system. After entering the correct user name and password, the administrator logs in to the main interface of the system. The interface mainly includes information collection module, face training module, face attendance module and attendance information query module.

2) Information collection module: The information collection module mainly completes the basic personal information and face information input of employees, including the basic personal information of the employee's name, number, phone number, gender, department, ID number and so on. Face information input includes two ways, one way is to call the camera to shoot 4 face photos, if the photo quality is qualified, saved in JPG format under the specified folder. If the photos do not meet the requirements, the staff will be prompted to take photos again; Another way is that employees can take good photos by themselves and send them to the administrator's email. The administrator can upload the photos directly on the computer.

3) Image preprocessing module: The image preprocessing module mainly processes the collected face images,because the port working environment is special,the face may be contaminated with dust and other impurities,causing a certain degree of influence on face recognition.In order to improve the accuracy of face recognition,the face is preprocessed before face training to reduce image noise pollution and enhance image features.And the processed photos were rotated, translated,mirrored and other operations to prepare for face training.

4) Face training module: After the pretreatment of face image,face training is needed.The system uses face detection algorithm RetinaFace and face recognition part uses an additive angular edge loss (ArcFace) algorithm.However, the combination of Multi-task Cascaded Convolutional Neural Networks(MTCNN) face detection algorithm and Facenet face recognition algorithm is better in the actual test environment.

5) Face attendance module: The face attendance module mainly completes the real-time attendance of employees. When the camera captures the face, it will collect the captured face and extract the corresponding face features, and then compare the collected face features with the face feature template in the database. If it is the face in the database, the second step is to judge the background environment taken by the staff when they punch in. If there are special characteristics of the port environment, the staff will be prompted to sign in successfully according to the location provided by the positioning system, and the sign in information will be written into the database. If it does not exist, prompt the staff to take a new picture and submit it; If the face information in the database cannot be matched, the employee will be reminded that the face cannot be matched.

6) Attendance information query module: The attendance information query module mainly completes the attendance information query of the employees.After the successful attendance,the attendance information of the employees is recorded,which mainly includes the employee's work number,name,check-in time,check-out time and attendance status.

SUMMARIZE

In this paper,the algorithms based on RetinaFace face detection and ArcFace face recognition are applied to port staff attendance in unrestricted scenarios.In order to obtain a more stable and accurate face representation method,this paper studies face detection,face image pretreatment, face recognition,loss function and other aspects.Through verifying ArcFace loss function has a good performance on PubFig data set.Using the camera to obtain face data information and face database can accurately identify whether the person is in the database.The feasibility of the model in port unrestricted scenarios is verified.

In the future,problems related to employee punch location,recognition rate,processing capacity and hardware equipment need to be further explored and realized.

ACKNOWLEDEGMENTS

The work described in this paper was supported by the Research Project of Wuhan University of Technology Chongqing Research Institute (YF2021-10).

REFERENCES

[1] LIU WY,WEN YD,YU ZD,et al.SphereFace:Deep Hypersphere Embedding for Face Recognition [C]//2017 IEEEConference on Computer Vision and Pattern Recognition(CVPR)IEEE,2017:6738-6746.

[2] WANG H,WANG YT,ZHOU Z,et al.Cosface:large margin cosine loss for deep facerecognition [C]//2018 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)IEEE,018:5265-5274.

[3] DENG JK,GUO J,XUE NN,et al.ArcFace:additive angular margin loss for deep face recognition [C]//2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)IEEE,2019:4690-4669

[4] Wang Mei,Zhang Yaobin,Deng Weihong. Meta Balanced Network for Fair Face Recognition.[J]. IEEE transactions on pattern analysis and machine intelligence,2021,PP:

[5] Qiu Haibo,Gong Dihong,Li Zhifeng,Liu Wei,Tao Dacheng. End2End Occluded Face Recognition by Masking Corrupted Features.[J]. IEEE transactions on pattern analysis and machine intelligence,2021,PP:

[6] Deng Jiankang,Guo Jia,Yang Jing,Xue Niannan,Cotsia Irene,Zafeiriou Stefanos P. ArcFace: Additive Angular Margin Loss for Deep Face Recognition.[J]. IEEE transactions on pattern analysis and machine intelligence,2021,PP:

[7] Liu Decheng,Gao Xinbo,Peng Chunlei,Wang Nannan,Li Jie. Heterogeneous Face Interpretable Disentangled Representation for Joint Face Recognition and Synthesis.[J]. IEEE transactions on neural networks and learning systems,2021,PP:

[8] Ming Xiang,Wei Fangyun,Zhang Ting,Chen Dong,Zheng Nanning,Wen Fang. Group Sampling for Scale Invariant Face Detection.[J]. IEEE transactions on pattern analysis and machine intelligence,2020,PP:

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

DP-Face: Privacy-Preserving Face Recognition Using Siamese Network

Nazhao Yan, Hang Cheng*, Meiqing Wang
School of Mathematics and Statistics
Fuzhou University
Fuzhou Fujian, China
Email: zoeyyannazhao@gmail.com, hcheng@fzu.edu.cn, mqwang@fzu.edu.cn

Qinjian Huang, Fei Chen
School of Computer Science and Big Data
Fuzhou University
Fuzhou Fujian, China
Email: 821403552@qq.com,chenfei314@fzu.edu.cn

***Abstract*—With the rapid development of deep learning, face recognition technology based on deep learning has been widely developed in recent years. However, during the training of the deep learning model, there is a risk of privacy leakage. If an attacker obtains private data, such as tags of the training data, the face images may be restored, and private information is leaked. To protect the private information of the face recognition model, we introduce differential privacy technology to propose a privacy-preserving face recognition scheme using the Siamese Network framework called DP-Face. Unlike other privacy-preserving face recognition methods, we can adjust the balance between privacy and utility through privacy budgets according to actual needs. Experimental results show that the effectiveness and privacy of the proposed DP-Face can be well guaranteed.**

***Keywords*-privacy-preserving; differential privacy; face recognition; deep learning**

I. Introduction

In recent years, following dramatic advances in artificial intelligence technologies, biometric recognition technology is rapidly integrating into people's lives. As one of the most popular biometric technologies, face recognition can identify individuals without their knowledge. In 2015, the FaceNet constructed by Schroff et al. achieved a recognition rate of 99.6%, which implies the reliability and practicability of face recognition technology [1]. Besides, Alipay provides facial payment services to more than 243 million users in China in 2020.

However, more and more security problems have gradually been exposed as deep learning is widely used in face recognition. Generally speaking, the training process of a deep learning model requires a large amount of representative data, which may contain sensitive information of the data owner such as age and gender. For the sake of security, the model should not disclose such private sensitive information. In a recent work [2], the attacker only needs to obtain the label of the data in the training set and get access to the model, and then the face image can be recovered from the model. In this case, the capability of effective face recognition is desired when keeping users' privacy confidential to the attacker. Obviously, it can be solved by employing the signal processing technologies in the encrypted domain [3], [4].

In this paper, we employ differential privacy technology to present a privacy-preserving face recognition (DP-Face). This scheme assumes that only one party holds sensitive private data, and we mainly focus on privacy leakage caused by model output. Our DP-Face uses Siamese Network to construct a similarity measurement network for better recognition accuracy. Meanwhile, to protect the privacy of training data, we introduce the differential privacy strategy. In addition, DP-Face can achieve the trade-off between the privacy and utility of the model by adjusting the privacy budget. The experiments illustrate that our DP-Face can well finish the task of face recognition while protecting private data from being leaked.

II. Preliminary

In this section, we review the basic concept of differential privacy technology and Siamese Network, which serve as the cornerstone for constructing the proposed DP-Face.

A. Differential Privacy

Differential privacy [5] ensures that inserting or deleting records in the data set will not affect the output results. And also, the protection model does not care about the background knowledge of the attacker. Even if the attacker has all records except a certain record, it cannot infer the undisclosed record. The formal definition of differential privacy is given below.

Definition 1 ((ε, δ)-DP). A randomized algorithm $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$ satisfies (ε, δ)-differential privacy if for any neighbouring datasets $d, d' \in \mathcal{D}$ and for any all output sets $S \subseteq \mathcal{R}$ it holds that

$$Pr[\mathcal{M}(d) \in S] \leq e^{\varepsilon} Pr[\mathcal{M}(d') \in S] + \delta, \qquad (1)$$

where d and d' have at most one record difference. Note that the randomness of the algorithm $\mathcal{M}$ means that for a specific input, the output of the algorithm is not a fixed value but obeys a certain distribution. The parameter ε represents the privacy budget. And the smaller the ε, the higher the degree of privacy guarantee.

The typical mechanisms for achieving differential privacy are the Laplace mechanism and Gaussian mechanism [6]. Among them, the Laplace mechanism provides a strict $(\varepsilon, 0)$-DP, and the Gaussian mechanism provides a relaxed (ε, δ)-DP [7]. In DP-Face, we use the Gaussian mechanism, which is defined as follows:

$$\mathcal{M}(x) \triangleq f(x) + N\left(0, \sigma^2 s_f^2\right), \qquad (2)$$

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00030

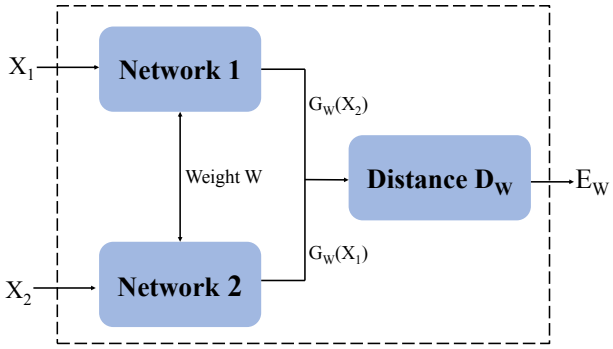


Figure 1. Architecture of Siamese Network.

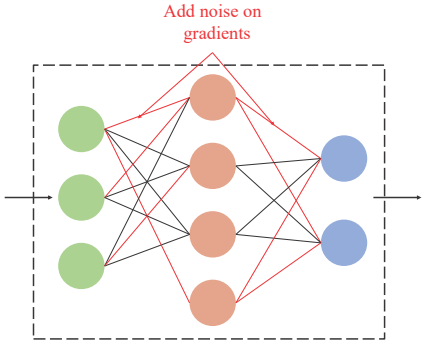


Figure 2. Construction of the backbone network of Siamese Network.

where s_f indicates the sensitivity related to the function f.

Definition 2 (l_2**-Sensitivity).** For any neighbouring datasets $d, d' \in \mathcal{D}$, given a function $f : \mathcal{D} \to \mathcal{R}$, the l_2-sensitivity Δf is defined as:

$$\Delta f = \max_{d,d'} ||f(d) - f(d')||_2. \quad (3)$$

B. Siamese Network

The goal of the Siamese Network [8], [9] is to calculate the similarity of two similar images. It has two identical sub-networks, both of which have the same parameters and weights. The special characteristic of this network is that its training samples take image pairs as input, and the features are extracted by two sub-networks respectively, and finally, the feature vector pairs of the samples are obtained. The architecture of the Siamese Network is shown in Figure 1.

Here, (X_1, X_2) represents the input image pair, $(G_W(X_1), G_W(X_2))$ is the output feature pair of network 1 and 2. Then, the similarity E_W of the sample pair will be predicted through a reasonable similarity measure.

In Siamese Network, the commonly used loss function is the contrastive loss function, which is defined as:

$$L(W) = \frac{1}{2}(1-Y)D_W^2 + \frac{1}{2}Ymax(0, m - D_W)^2, \quad (4)$$

where $D_W = \|G_W(X_1) - G_W(X_2)\|_2$ denotes the Euclidean distance between the outputs, Y denotes the label, and m is the marginal value greater than 0.

III. Privacy-Preserving Face Recognition Using Siamese Network

In this section, we present the proposed Privacy-Preserving Face Recognition with Siamese Network (DP-Face) model and the privacy Analysis of DP-Face.

A. DP-Face Framework

The Siamese Network in our DP-Face scheme is designed on a pre-trained network. Pre-trained VGG16, Inception, or ResNet in ImageNet is often selected as a candidate backbone network for Siamese Network. Meanwhile, The output of the fully connected layer of the backbone network is used as a feature of the input image.

Note that the gradient contains private information about the dataset in the deep learning model. It can be ensured that the subsequent updating operation of the parameter value will not leak user information as long as the gradient is disturbed. Therefore, instead of adding noise to the final parameters, we add noise proportional to the training data on the gradient of the Wasserstein distance. The specific details of DP-Face are shown in Figure 2 and Algorithm 1.

Algorithm 1: Differentially private Stochastic Gradient Descent in DP-Face

Input: Examples $x_1, x_2, \ldots, x_N$, learning rate α_t, number of iterations n, group size L, gradient norm bound C, loss function $\mathcal{L}(\theta) = \frac{1}{N}\mathcal{L}(\theta, x_i)$.

Output: θ_n and compute the overall privacy cost (ε, δ) using a privacy accounting method.

Initialize model parameters θ_0

for $t = 1$ *to* n **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $g_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\overline{g}_t(x_i) \leftarrow g_t(x_i)/max(1, \frac{\|g_t(x_i)\|_2}{C})$

Add noise

$\widetilde{g_t} \leftarrow \frac{1}{L}(\sum_i \overline{g}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}))$

Descent

$\theta_{t+1} \leftarrow \theta_t - \alpha_t \widetilde{g_t}$

end

B. Privacy Analysis of DP-Face

To prove that DP-Face satisfies differential privacy, we analyze the privacy loss of our DP-Face model.

Definition 3 (Privacy Loss). For the given neighbouring datasets d and d', assuming aux is the auxiliary input and $o(o \in \mathcal{R})$ is the output, the privacy loss at o can be defined as:

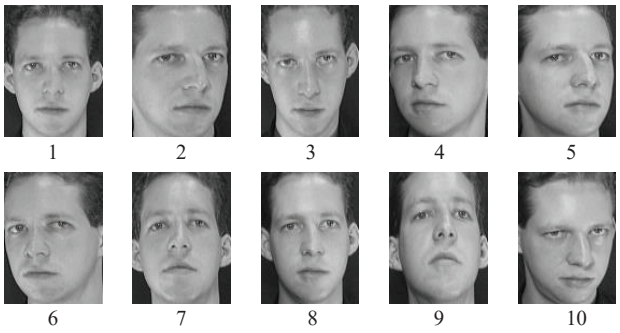


Figure 3. Sample images from ORL Database of Faces.

$$c(o;\mathcal{M},aux,d,d') \triangleq log\frac{Pr[\mathcal{M}(aux,d)=o]}{Pr[\mathcal{M}(aux,d')=o]} \quad (5)$$

The privacy loss is introduced to describe the distribution difference between two data. Besides, the privacy loss random variable $c(o;\mathcal{M},aux,d,d')$ is used to describe the privacy budget of $\mathcal{M}(d)$ in **Definition 1**.

Definition 4 (Log moment generation function). For the randomized algorithm $\mathcal{M}$, define the λ-th moment $\beta_{\mathcal{M}}(\lambda;aux,d,d')$ as the log of moment generation function evaluated at λ:

$$\beta_{\mathcal{M}}(\lambda;aux,d,d') \triangleq \log \mathbb{E}_{o\sim\mathcal{M}(aux,d)}\left[\exp\left(\lambda c\left(o;\mathcal{M},aux,d,d'\right)\right)\right]. \quad (6)$$

Definition 5 (Moment Accountant). The moment accountant is defined as:

$$\beta_{\mathcal{M}}(\lambda) \triangleq \max_{aux,d,d'} \beta_{\mathcal{M}}(\lambda;aux,d,d'), \quad (7)$$

In our DP-Face, we use the moment accountant to track the privacy loss caused by the published model and provide a privacy loss boundary with high accuracy.

The following theorems and lemma can be proved in the works [5], [10], which ensure our DP-Face scheme satisfies (ε,δ)-DP guarantee.

Theorem 1 (Composability). Suppose that a randomized algorithm $\mathcal{M}$ consists of a sequence of adaptive mechanisms $\mathcal{M}_1,\cdots,\mathcal{M}_k$, where $\mathcal{M}_i:\prod_{j=1}^{i-1}\mathcal{R}_j\times\mathcal{D}\rightarrow\mathcal{R}_i$. For any λ and $\alpha_{\mathcal{M}}(\lambda)$,

$$\alpha_{\mathcal{M}}(\lambda)\leq\sum_{i=1}^{k}\alpha_{\mathcal{M}_i(\lambda)}.$$

Theorem 2 (Tail bound). For any $\varepsilon>0$, the randomized algorithm $\mathcal{M}$ is (ε,δ)-differentially private for

$$\delta=\min_{\lambda} exp(\alpha_{\mathcal{M}}(\lambda)-\lambda\varepsilon).$$

Lemma 1. For any $\delta\in(0,1)$, $\sigma>\frac{\sqrt{2\ln(1.25/\delta)}\Delta f}{\varepsilon}$, the noise $Y\sim N(0,\delta^2)$ satisfies (ε,δ)-DP.

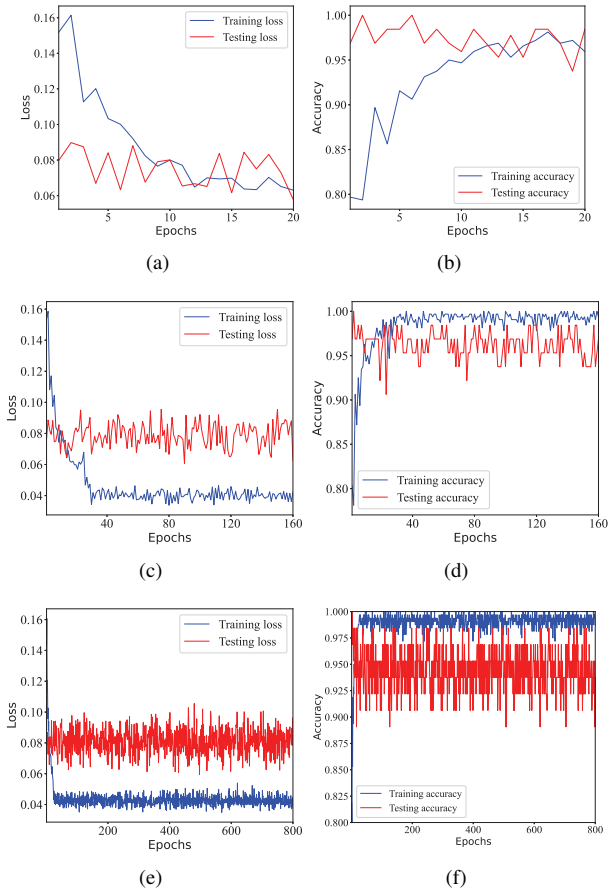


Figure 4. The loss and accuracy of DP-Face at different noise levels: (a) Loss at large noise level; (b) Accuracy at large noise level; (c) Loss at medium noise level; (d) Accuracy at medium noise level; (e) Loss at small noise level; (f) Accuracy at small noise level.

IV. Performance of DP-Face

To evaluate the performance of DP-Face, we utilize Tensorflow and Keras in Python 3 on one server, which is equipped with an 8-core Intel Core i7-9700 CPU @3.00GHz and 8GB of RAM running Windows 10-64bit. All the experiments were conducted with the ORL database created by Olivetti Research Laboratory in Cambridge, England.It contains 40 directories, each of which represents a different person, and contains ten face images. All the images are in PGM format, and each image is sized at 92×112, with 256-level grey per pixel. The partial face images are displayed in Figure 3.

A. Data Processing

The images in the ORL dataset are divided into the training set and the test set, where the training set is the first 37 directories of the 40 directories in the ORL and the last three directories for the test set. Considering the training set (370 face images in total) is limited, we use the pre-trained ResNet50 as the backbone network of the Siamese Network for feature extraction.

B. Performance Evaluation

In the plaintext model corresponding to our DP-Face, for the output of ResNet50, we first calculate the Euclidean

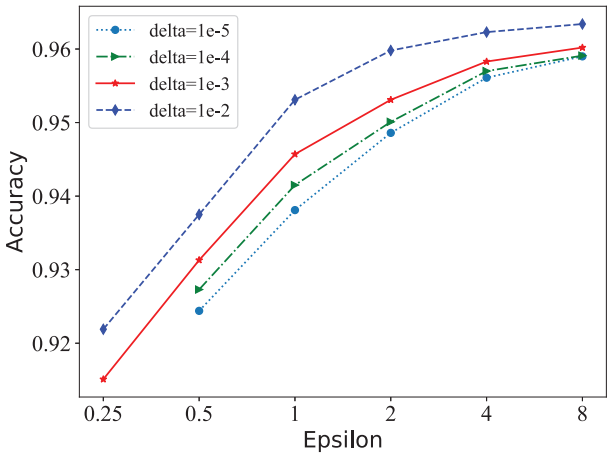


Figure 5. Effects of different parameters on classification accuracy.

distance between two output feature, then use the Dense layer and the Activation layer to construct the similarity calculation network. Here, the activation function is ReLU and the loss function is contrastive loss function. Using the batchsize with 64, we can reach the accuracy of 98.44% in the plaintext domain.

To better observe the effect of DP-Face, we conducted the experiments to evaluate the loss and accuracy of the results under different noise levels. Specifically, the noise is set to $0.05, 0.10, 0.20$, respectively. The higher the value, the more noise is added. Figure 4 shows the training and testing loss/accuracy as a function of the number of epochs in each plot. The toss accuracy performance varies with respect to the levels of the noises. As the noise increases, the accuracy of DP-Face will decrease, whereas the more the loss is generated. In the case of high noise, medium noise, and small noise, the accuracy was 95.31%, 96.87%, and 98.40%, respectively. In other words, when the added noise range is $0.20 \sim 0.05$, the accuracy of DP-Face remains above 95%, which means our DP-Face can complete face recognition tasks with high accuracy.

Besides, we verify the effect of differential privacy budget ε and a fixed noise scale δ for accuracy of DP-Face with the Gaussian mechanism. Here, privacy budget ε ranges from 0.1 to 10 and the noise scale δ is $10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$, respectively. From Figure 5, we can obtain the accuracy of the different (ε, δ), for example, when $\varepsilon = 0.25$ and $\delta = 0.01$, the accuracy of our DP-Face is 92.19%. In addition, it can be observed that for a fixed privacy budget ε, changing the δ value has a small impact on accuracy, but for a fixed noise scale δ value, changing the ε value has a large impact on accuracy. Plus, for a fixed δ, when more noise is added, the intensity of privacy protection will be greater; that is, the privacy budget ε will be smaller. However, it is not hard to find from Figure 5 that when the privacy budget is smaller, the model accuracy is lower. It means that the availability of the face recognition scheme will also decrease. Nevertheless, for the above (ε, δ), the accuracy of our DP-Face remains above 90%. Therefore, this framework can ensure that the accuracy of face recognition remains at a high level while protecting the privacy of face images.

V. Conclusion

In this paper, differential privacy technology and Siamese Network are used to construct a privacy-preserving face recognition scheme, which preserves the privacy of face training data in a differentially private case. The DP-Face model mitigates information leakage by adding designed noise to the gradients during the deep learning process. In addition, experiments show that the DP-Face can still converge under the constraints of noise-added training data and achieve high-accuracy face recognition.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 62172098), the Natural Science Foundation of Fujian Province (No. 2020J01497).

References

[1] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[2] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[3] H. Cheng, X. Liu, H. Wang, Y. Fang, M. Wang, and X. Zhao, "Securead: A secure video anomaly detection framework on convolutional neural network in edge computing environment," *IEEE Transactions on Cloud Computing*, 2020.

[4] H. Cheng, H. Wang, X. Liu, Y. Fang, M. Wang, and X. Zhang, "Person re-identification over encrypted outsourced surveillance videos," *IEEE Transactions on Dependable and Secure Computing*, 2019.

[5] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[6] J. Dong, A. Roth, and W. J. Su, "Gaussian differential privacy," *arXiv preprint arXiv:1905.02383*, 2019.

[7] Z. Bu, J. Dong, Q. Long, and W. J. Su, "Deep learning with gaussian differential privacy," *Harvard data science review*, vol. 2020, no. 23, 2020.

[8] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546.

[9] I. Melekhov, J. Kannala, and E. Rahtu, "Siamese network features for image matching," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 378–383.

[10] F. Fioretto and P. Van Hentenryck, "Privacy-preserving federated data sharing." in *AAMAS*, 2019, pp. 638–646.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

SSD small object detection algorithm based on feature enhancement and sample selection

Zhipeng Liu
School of Artificial Intelligence and Computer
Jiangnan University
Wuxi, China
952738565@qq.com

Wei Fang
School of Artificial Intelligence and Computer
Jiangnan University
Wuxi, China
fangwei@jiangnan.edu.cn

Jun Sun
School of Artificial Intelligence and Computer
Jiangnan University
Wuxi, China
junsun@jiangnan.edu.cn

***Abstract*—SSD has a poor detection effect on small objects. The first reasons is its insufficient feature extraction for small objects. To solve this problem, a feature enhancement module is proposed to make better use of the information around the object to improve the identification ability of small objects. The second reason is that the division of positive and negative samples is unreasonable. The threshold is unfriendly for small objects. To solve this problem, an adaptive training sample selection algorithm is adopted to select the threshold. To improve SSD by the above two methods, and experiments on the PASCAL VOC data set. The mean accuracy precision is increased by 2.6% compared to the SSD algorithm. Compared with the series of SSD improved algorithms such as DSOD, RSSD, DSSD, FSSD, the mAP of our method increased by 2.1%, 1.3%, 1.2%, 1.0%. Our method significantly improved the detection effect of small objects, surpassing SSD and its improved algorithms.**

Keywords-object detection; SSD; small object; feature enhancement; sample selection

I. INTRODUCTION

object detection [1] is one of the important tasks of computer vision. It contains two sub-tasks, one is to accurately identify the object, and the other is to accurately locate the object. In recent years, with the rapid development of deep learning [2], neural networks can extract object features well. Object detection methods based on convolutional neural networks have become a hot issue in the field of object detection. The feature obtained through the deep learning method is stronger than the traditional method. In the wave of deep learning, two types of object detection algorithms have emerged. The one-stage detection algorithms are represented by SSD [3], DSSD [4], DSOD [5], RetinaNet [6], YOLO [7] series, etc. The two-stage detection algorithms are represented by Faster R-CNN [8], Cascade R-CNN [9] and so on. The one-stage detection algorithm has a speed advantage, and the two-stage detection algorithm has a precision advantage, and both have their own advantages.

Small object detection is a difficult task of object detection. The small object occupies few pixels, its image resolution is low, the information is insufficient, and there are fewer features for learning, which leads to the model's poor feature expression ability for small objects. To solve this problem from the perspective of scale, there are FPN [10] algorithm, SNIP [11] algorithm, etc, which integrate high-level feature map with low-level feature map, and make full use of the semantic information of high-level features and the resolution feature of low-level features to improve the expression ability of low-level features; To solve this problem from the perspective of learning the surrounding information of small objects, such as SSH [12,13], etc, which increases the receptive field of the convolutional layer, and better obtain the small object's information through context information information to improve the ability of expressing small objects; To solve this problem from the perspective of the default box, representative of S3FD [14], etc. The author has done a detailed default box experiment, and designed the default box more reasonably, so the default box can match small objects better. To solve this problem from the perspective of the matching strategy, such as Cascade R-CNN, etc, which do not set too strict IoU threshold to ensure the number of small objects' default boxes. The above-mentioned solutions have improved the detection effect of small objects, which has greatly inspired the research work of this paper.

In order to improve the detection effect of SSD on small objects, this paper starts from two perspectives. Firstly, this paper proposes a feature enhancement module, which can better learn the surrounding information of small objects and make up for the shortcomings of insufficient features of small objects; Secondly, this paper adopts a positive and negative sample selection strategy. The original SSD's determination threshold for positive samples is too strict. The number of default boxs corresponding to the object is originally relatively small. After the hard threshold filtering, the number of positive sample default boxs corresponding to the remaining small objects will be less, resulting in insufficient training of small objects, so this paper adopts an adaptive training sample selection method. Choose an appropriate threshold for each object, and use this threshold for sample selection. The experimental results show that the improved SSD algorithm in this paper has significantly improved the detection effect of small objects. From the results of the PASCAL VOC2007 test set, the detection accuracy of the three categories such as bottle, pottedplant and chair has been significantly improved. The three categories have a large number of small objects. Our method is better than SSD and a series of SSD improved algorithms.

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00031

II. SMALL OBJECT DETECTION ALGORITHM

This paper improves SSD from two aspects. Firstly, a feature enhancement module is proposed to supplement the feature of the small object by fully learning the edge information of the small object, and improve the detection ability of the SSD detector for the small object. Secondly, adopt an adaptive positive and negative sample selection strategy to replace SSD's original strategy. The strategy adaptively select the IoU threshold to determine the positive which can ensure the number of positive sample boxes for small objects. The improved SSD algorithm proposed in this paper significantly improves the detection effect of small objects.

A. Feature Enhancement Module

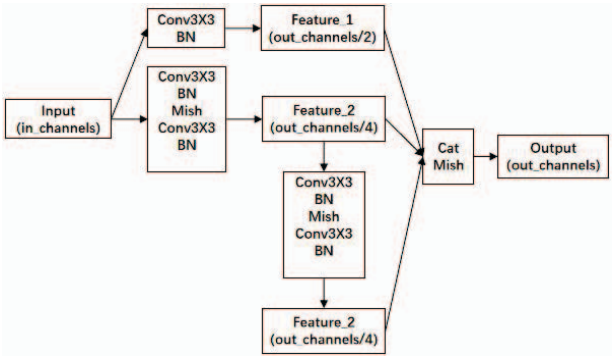


Figure 1. Network structure of feature enhancement module.

Feature enhancement module, we can called it FM. Fig. 1 shows its structure.

The size of the convolution kernel used by the FM is 3. If the convolution kernel is 1, rich surrounding information cannot be extracted. If the size of the convolution kernel is larger, it will increase a lot of parameters and increase the computational cost. After convolution, the FM uses BatchNorm to normalize the data, adjust the data distribution, and accelerate training. The FM module uses the Mish function as the activation function. In view of the long-term dominant position of ReLU in the activation function of deep learning, we compares the Mish function and the ReLU function. The formulas of Mish and ReLU are respectively as follows.

$$g(z) = z * \tanh\left(z * \ln\left(1 + e^{z}\right)\right)$$

$$g(z) = \begin{cases} z & z > 0 \\ 0 & z \le 0 \end{cases}$$

The function image of the two function is shown in Fig. 2. It is not difficult to find that ReLU is directly set to zero for negative values, and Mish [15] has a better gradient flow for negative values, so this module uses Mish as the activation function.

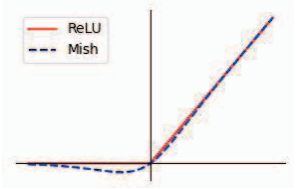


Figure 2. Image function of ReLU and Mish.

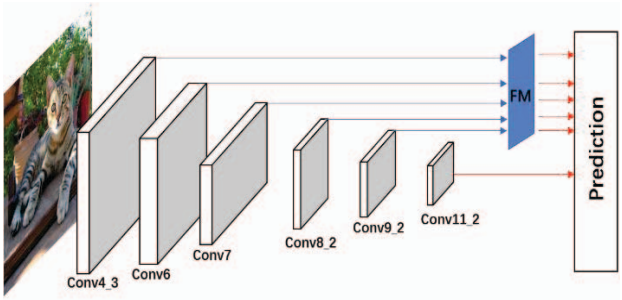


Figure 3. Network structure of SSD embedded FM.

The structure of FM embedded in the original SSD is shown in Fig. 3. In this paper, after the feature map of Conv4_3, Conv6, Conv7, Conv8_2, Conv9_2, the FM module is connected. Then put the five layer's output and Conv11_2 feature map into prediction branch. We only uses the FM module for the first five layers of feature maps, and the last layer of feature maps does not use the FM module. The reason is that the size of the last layer of feature maps Conv11_2 has been 1, and the kernel size 3 in the FM module is no longer applicable.

B. Adaptive Training Sample Selective

The adaptive training sample selection [15] method can better determine the IoU threshold. It selects positive and negative samples according to the statistical characteristics of the ground truth boxes. The IoU threshold can be determined adaptively for each real object. This method doesn't increase any computational overhead to improve the performance of SSD detector. The algorithm steps are as follows.

Input: G、L、A_i、A、k .Output: P、N . G represents the set of all ground truth boxes in the image; L represents the number of feature maps; A_i is the set of default boxes for the i-th feature map; A is the set of all default boxes; k represents the number of default boxes selected from each feature map; P represents the set of positive samples; N represents the set of negative samples.

a) For each ground truth g , $C_g \leftarrow \Phi$.

b) For each feature map i , $i \in [1, L]$,select k default boxes whose center closest to g from A_i based on L2 distance. These selected boxes as S , $C_g = C_g \cup S$.

c) Compute IoU between C_g and g , as D_g .

d) Compute mean of D_g , as m_g .

e) Compute variance of m_g , as v_g .

f) For each g , set threshold as t_g , $t_g = m_g + v_g$.

g) Compute mean of D_g , as m_g .For each candidate box o , $o \in C_g$, if IoU between o and g larger than t_g and its center in g , it will be viewed as positive sample, written as P , otherwise as negative sample, written as N , $N = A - P$.

h) Return P、N .

The algorithm is not sensitive to the value k . In this paper, experiments are conducted on the PASCAl VOC dataset, and the final value selected is 11, because the results obtained at this time are slightly better than other values.

III. EXPERIMENT

A. Experienmental Steps

a) Build this paper's network.

b) Train VGG16 on the ImageNet, get pretrained VGG16 .

c) Make migration learning, remove the fully connected layer of the VGG16 model, load its weights as the initial parameters of the backbone.Use kaiming distribution to initialize the category prediction branch and location prediction branch, and initialize the rest of the parameters randomly.

d) Load the VOC2007 and VOC2012 training set pictures, set the picture size to 300×300, use random cropping, flipping and other methods for data augmentation, and input the network for training.

e) Use stochastic gradient descent method as the optimizer, the initial learning rate is set to 0.001, the batchsize is set to 32, and the maximum number of iterations is set to 200.

f) Save the training model.

B. This paper's algorithm compared with original SSD

As shown in Table. 1. Compared with the original SSD algorithm, the algorithm in this paper has significantly improved the detection effect on bottle, potted plant, and chair which have a large number of small objects. AP increased by 10.8%, 4.0%, and 2.4% respectively. These three categories' mAP increased by 5.7%, and the rest 17 categories' mAP increased by 2.1%. The mAP of the overall twenty categories of objects increased from 77.2% to 79.8%.

TABLE I. COMPARISION OF OURS AND SSD ON ALL CLASSES

	SSD300/%	**ours/%**
mAP	77.2	79.8
aero	82.2	86.4
bike	84.7	87.9
bird	74	78.8
boat	68.8	73.9
bottle	50.1	60.9
bus	84.4	87.5
car	86.2	86.8
cat	87.9	86.6
chair	61.7	64.1
cow	82.0	84.0
table	74.5	81.1
dog	85.4	86.5
horse	87.2	85.8
mbike	83.2	85.2
person	78.1	79.6
plant	51.3	55.3
sheep	77.5	79.4
sofa	80.7	80.4
train	87.7	85.8
tv	76.6	79.7

Fig. 4 is a comparison chart of the recall and precision of the above three categories.The solid line represents the algorithm in this paper, and the dotted line represents the original SSD algorithm. It can be seen that the algorithm in this paper has a significant improvement in the recall rate of small objects.

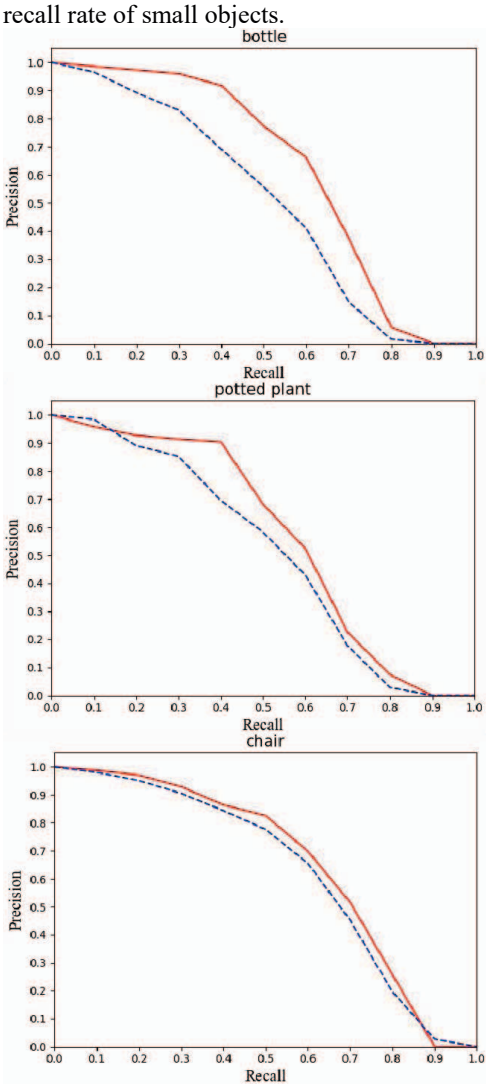


Figure 4. Comparison of recall and precision.

C. This paper's algorithm compared with SSD's improved algorithm

TABLE II. COMPARISION OF OUR METHOD AND OTHER METHOD BASED ON SSD

Method	**Backbone**	**Input size**	**Boxes' num**	**Speed (FPS) Titan X**	**mAP**
SSD300	VGG	300×300	8732	46	77.2
DSOD300	VGG	300×300	---	17.4	77.7
RSSD300	DS/ 64-192-48-1	300×300	8732	35	78.5
DSSD321	ResNet-101	321×321	17080	9.5	78.6
FSSD300	VGG	300×300	8732	35.6	78.8
ours	VGG	300×300	8732	38.3	79.8

As shown in Table. 2, compared with the SSD's improved algorithms DSOD, RSSD, DSSD, FSSD, the algorithm in this paper has the lowest input image resolution, the smallest number of default boxes, the fastest speed, and the highest accuracy.

The intuitive comparison of speed and accuracy is shown in Fig. 5.

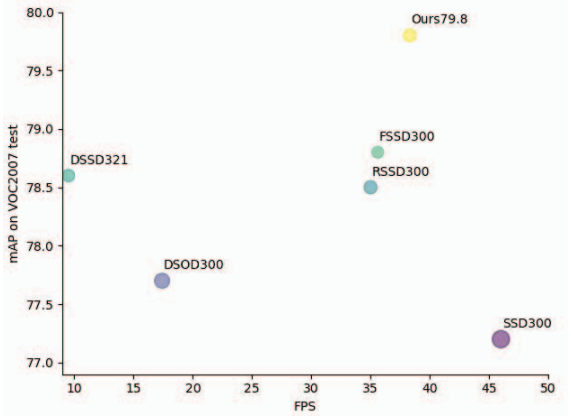


Figure 5. Comparison of speed and mAP.

D. Pictures shows

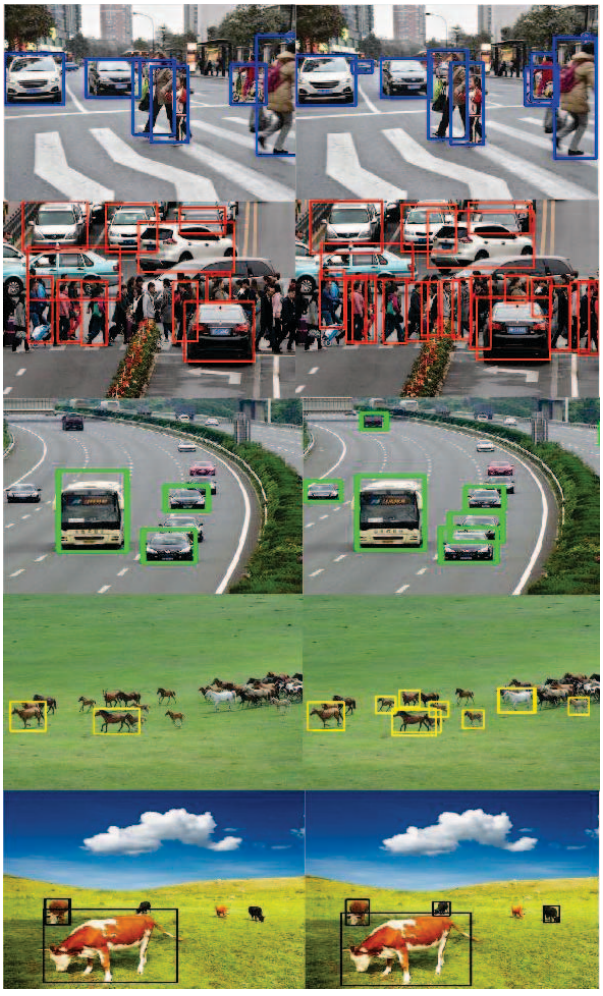


Figure 6. Picture shows of our method and SSD.

In order to verify the effectiveness of the improved SSD algorithm in this paper, we download some pictures with small objects from the Internet and compare the experimental results with the original SSD algorithm. As shown in Fig. 6, the labels and scores are removed for better observation. For the same picture, the left one is the detection result of the original SSD algorithm and the right one is the detection result of the improved SSD algorithm in this paper.

IV. CONCLUSION

The results of comparative experiments show that compared with the original SSD algorithm, the improved algorithm in this paper has significantly improved the detection effect on the VOC2007 test set, and surpasses a series of improved algorithms such as DSSD, DSOD, and FSSD in accuracy and speed. This proves the efficiency and advantages of the algorithm in this paper.

REFERENCES

[1] Wu X, Sahoo D, Hoi S C H. Recent advances in deep learning for object detection[J]. Neurocomputing, 2020.

[2] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798-1828.

[3] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[4] Deconvolutional Single Shot Detector[J]. 2017.Fu C Y, Liu W, Ranga A, et al. Dssd: Deconvolutional single shot detector[J]. arXiv preprint arXiv:1701.06659, 2017.

[5] Shen Z, Liu Z, Li J, et al. Dsod: Learning deeply supervised object detectors from scratch[C]//Proceedings of the IEEE international conference on computer vision. 2017: 1919-1927.

[6] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[8] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.

[9] CaiCai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition.2018:6154-6162

[10] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[11] Singh B, Davis L S. An analysis of scale invariance in object detection snip[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3578-3587.

[12] Najibi M, Samangouei P, Chellappa R, et al. Ssh: Single stage headless face detector[C]//Proceedings of the IEEE international conference on computer vision. 2017: 4875-4884.

[13] Deng J, Guo J, Zhou Y, et al. Retinaface: Single-stage dense face localisation in the wild[J]. arXiv preprint arXiv:1905.00641, 2019.

[14] Zhang S, Zhu X, Lei Z, et al. S3fd: Single shot scale-invariant face detector[C]//Proceedings of the IEEE international conference on computer vision. 2017: 192-201.

[15] Misra D. Mish: A self regularized non-monotonic neural activation function[J]. arXiv preprint arXiv:1908.08681, 2019.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Application of CNN-based Face Recognition Technology in Smart Logistics System

Geng Miao
Wuxi Taihu University,Wuxi 214064,China
Jiangsu Key Lab of IoT Application Technology, Wuxi Taihu University,Wuxi 214064,China

***Abstract*—Smart logistics combines the Internet of Things with big data, cloud computing and other information processing technologies to improve all aspects of logistics business and optimize the status quo of logistics services. This research aims to study the application of Convolutional Neural Network-based face recognition technology in intelligent logistics systems. First, briefly describe the key concepts of Faster Region with CNN feature, convolutional neural networks, the essence of smart logistics, and the background of the overall architecture design of CNN-based smart face recognition logistics, including data modules, data storage modules, and data presentation layers. Analysis of bottlenecks in other smart logistics systems, such as time-consuming large-capacity storage. By choosing WIDER Face as the sample training set and FDDB as the sample test set, the traditional face detection algorithms are compared. Experimental results show that CNN has the fastest detection speed. And the loss rate of CNN is low, and the accuracy can reach 91% under the condition of unlimited faces.**

Keywords-Convolutional Neural Network; Face Recognitio; Smart Logistics; Logistics System

I. INTRODUCTION

With the rapid development of electronic information and Internet technology and in-depth network traffic increasing rapidly, data security is becoming more and more important [1]. Compared with traditional authentication methods, identity recognition is an important part of information security [2]. Biometric elements have many advantages, such as uniqueness, strong anti-counterfeiting in actual use, and strong memory [3]. Facial data is a natural way of communicating with each other and is usually easily accepted by the public. Compared with fingerprints and iris, collecting faces does not require the cooperation of the collector, and the collection process is more natural [4-5].

Based on the above advantages, face recognition has always been a research hotspot in academia, attracting the research interest of the majority of researchers, and the results are also remarkable. Lu J proposed a Face Recognition method that is easy to train and based on a single CNN. Their CNN model utilizes the remaining learning framework. In addition, it uses standardized features to calculate the loss. Their large number of experiments show that they have good generalization ability on different data sets. They have obtained very competitive and state-of-the-art results on LFW, IJB-A, YouTube faces and CACD datasets [6]. Peng C proposed a new heterogeneous face recognition (HFR) method based on graphical representation. The Markov network is used to represent the heterogeneous image blocks, and the spatial compatibility between adjacent image blocks is considered. A coupled representation similarity measure (CRSM) is designed to measure the similarity between graphical representations. Extensive experiments on multiple HFR scenes (view sketches, forensic sketches, near-infrared images, and thermal infrared images) show that the proposed method outperforms the state-of-the-art methods [7-8]. Therefore, the study of face recognition based on CNN has high practical significance.

The innovation of this paper is to use a convergent neural network to create a logistics approval system based on face recognition. At the same time, through face recognition, the misunderstandings and situations in the traditional logistics links are eliminated. Wrong downloads will reduce errors in the signing process. On the other hand, the information verification time in the signature process is shortened, and the efficiency of the signature is improved. This research can further improve the accuracy of face recognition in offline logistics, at the same time improve the efficiency of offline logistics, and further accelerate the information development of the logistics industry.

II. RESEARCH ON CNN-BASED FACE RECOGNITION TECHNOLOGY IN SMART LOGISTICS SYSTEM

A. Convolutional Neural Network

Convolutional neural network has become a research hotspot in the field of language and visual perception [9]. Convolutional neural networks are multilayer perceptrons, especially used to recognize two-dimensional graphics, such as conventional filters, but this network structure has strong resistance to tilt, scaling, deformation, or other forms of distortion. It can be customized to avoid the traditional process of extracting complex features and reconstructing data.

B. The Overall Architecture Design of CNN-based Face Recognition Smart Logistics

The essence of logistics is to serve customers and markets. The logistics industry is a complex service industry that has important strategic significance to the country's economic development. The comprehensive service functions of logistics enterprises and the nodes in the city improve the quality of logistics services, and it is of great significance to realize logistics modernization [10-11].

Before the development of the system, it is necessary to build the overall framework of the face recognition logistics system based on deep learning. The overall process of face recognition is designed in this way. The following mainly introduces the overall architecture design of the deep learning face recognition logistics system.

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00032

1) Data processing unit. The data processing unit is composed of an image processing server, and the image processing server has a built-in cohesive neural network model based on deep learning. The main function of the image processor is to use multiple images uploaded by the user to the image storage unit as a set of training data for the training of the face recognition model. The recipient video uploaded by the courier is processed as a test data set to predict the result of image sorting. When uploading the sorting result to the file module, the database storage unit uploads the sorting result to the data display layer.

2) Data storage unit. The data storage unit is composed of a file server and a database server. The file server mainly stores and saves the photo of the person uploaded by the recipient and the key frame filtering of the video uploaded by the courier, and uses them as a test data set. Next, sort these two types of photos. The pictures uploaded by users are used as the training data set for the face recognition of the convergent neural network. Use the image uploaded by the processor as the test chart of the cohesive neural network.

3) Data display layer. The main function of this layer is to display the facial recognition classification results based on the deep convergence neural network in express delivery. The courier compares the receipt of the goods based on the feedback information.

C. Logistics Signing Management System

When logging in to the management interface, you will see the user login option. You must enter the previously registered user account and then enter the password. The system background decides whether the user login or the administrator login is based on the permissions at the time of registration. If you forget the password you used when registering, you can click Forgot Password to get it.

The logistics sign-off system can be divided into two categories: shippers and users. Once the processor is connected to the interface, you can view the sender, order details for today's shipment, order number, content, and shipping address through the order details. If the goods have been delivered to the specified address, you can confirm that the recipient is the user. Then click Confirm Shipment.

After arriving at the designated receiving address, the courier must use the mobile phone to collect the recipient's face video, connect to the face recognition system, upload the collected video, and analyze it.

After connecting to the system, the courier selects the face recognition system, clicks the file selection button, and chooses to upload the video taken by the face. After selecting the file, click the face recognition button to perform face recognition. The system can basically perform face recognition in real time: below the photo, the user name and location of the person are displayed. The sender compares the returned user name with the recipient of the order details. If the comparison is successful, the sender will deliver the goods and click the order details page to complete the delivery.

III. Investigation and Research of CNN-based Face Recognition Technology in Smart Logistics System

A. Data Set

This article uses two sets of face-to-face data to train the network model. The two most reliable face data sets for face detection are the FDDB face data set and the WIDER face data set.

FDDB is a set of evaluation data created specifically for face detection and is maintained by the School of Computer Science at the University of Massachusetts. It is one of the world's leading face tracking and evaluation data sets. The total number of people is 2,375, and each photo has a detailed label on the face coordinates. These faces were taken in the field, very close to everyday application scenarios, such as rare poses, occlusion, low resolution, and out of focus.

WIDER Face is a large-scale face detection reference data released by the Chinese University of Hong Kong. Including more than 3,000 photos and nearly 40,000 faces. All these data sets are manually labeled and contain real scenes of various facial image changes, including occlusion, posture, stop, and lighting.

B. Preprocessing of Data Entry

In this article, we will choose WIDER Face as the sample practice set and FDDB as the sample test set. Build the network after practicing the candidate framework. The candidate frames created should be divided into positive samples and negative samples. In other words, take a face photo as a positive example. A photo without a human face serves as an example of the negative side. The separation between positive and negative samples should be based on the overlap ratio. It is a parameter of the target detection and evaluation system. The overlap ratio is the ratio of the sum and the intersection of the detection result and the Ground Truth box labeled in the training set. The fetching results here are displayed in the option box. The calculation is shown in Equation 1. The specific calculation is shown in formula 1:

$$IOU = \frac{Detectionr\,\mathrm{Re}\,sult \cap GroundTruth}{Detectionr\,\mathrm{Re}\,sult \cup GroundTruth} \quad (1)$$

Calculate the IOU values of all candidate boxes and Groud Truth Box, and use the candidate box with the largest IOU as a positive sample to ensure that each picture has at least one positive sample. In addition, the candidate box with an IOU greater than 0.7 is also used as a positive sample. Candidate boxes less than 0.3 are regarded as negative samples.

Use T to represent the positive sample, then T=TP+FN, and N to represent the negative sample, then N=FP+TN, and the accuracy rate is shown in formula 2:

$$Accuracy = \frac{TP + TN}{P + N} \quad (2)$$

IV. INVESTIGATION AND RESEARCH ANALYSIS OF CNN-BASED FACE RECOGNITION TECHNOLOGY IN SMART LOGISTICS SYSTEM

A. CNN's Face Recognition Technology Algorithm Detection Speed Comparison

This article compares CNN with several more classic algorithms, including Multi Task CNN and Joint Cascade's classic face detection methods, all of which are tested using the FDDB database evaluation method. In addition, the comparison of time and test results are shown in Figure 1.

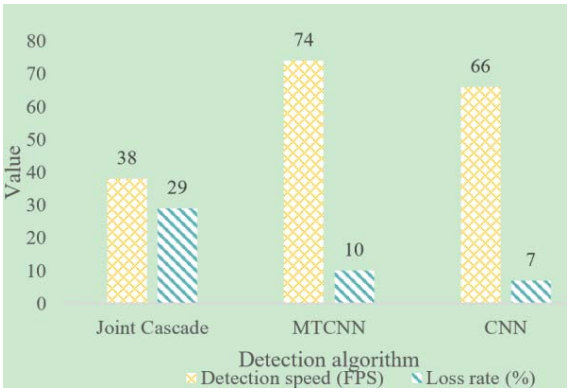


Figure 1. Algorithm comparison

The detection speed is expressed by FPS, that is, the number of frames transmitted per second. The higher the number of frames, the higher the algorithm detection speed. The improved CNN detection speed is uniformly controlled on the GPU (NVIDIA GTX1050). Compared with the three MTCNN methods Joint Cascade and Cascade CNN, CNN has the lowest loss rate. As shown in Figure 1, the detection speed is higher than Joint Cascade but lower than MTCNN, but it is sufficient to meet the needs of face-to-face systems.

B. Comparison of Correct Recognition Rate of CNN's Face Recognition Technology Algorithm

Because the database is an infinite choice of people, the recognition rate of some classic algorithms is not very high and cannot reach a satisfactory level. The detection rates of PCA (Principal Component Analysis), SVM (Support Vector Engine) and LBP (Local Binary Operation) are all lower than 50%. LBP can recognize effective faces under the condition of small rotation angle changes and uniform illumination, and it has certain robustness. However, the data range of FDDB exceeds the limit of LBP, so the recognition rate in the FDDB database is very low, less than 50%. Although SVM has the characteristic of approximating arbitrary functions, the characteristics of human faces in the database are complex and changeable. The general characteristics of PCA are not suitable for unlimited face environments, so the recognition rate is the lowest. The accuracy of science fiction (sparse filtering) is much higher than the first three algorithms, because it can dilute facial features to achieve optimized high-resolution features, and the recognition rate of global abstract features is just over 70%. Compared with the above DBN (Deep Belief Network), the recognition accuracy is very high, so advanced features can be removed.

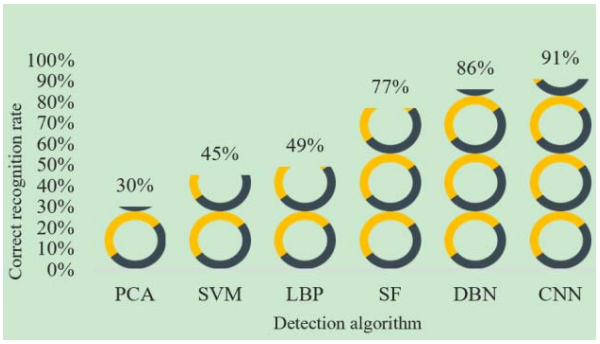


Figure 2. Recognition rate of various algorithms in LFW database

Comparing CNN with the above algorithm, as shown in Figure 2, it can be seen that CNN has the highest recognition rate. The recognition of DBN is also very high, but the loss of CNN is lower. Under infinite face conditions, an accuracy of 91% can be obtained, which shows that CNN can be applied to a face recognition system to achieve the desired effect.

V. CONCLUSIONS

As the most popular recognition technology at present, face recognition has a very wide application potential in real world applications due to its non-contact, convenient and other advantages. Especially in recent years, due to the rapid development of artificial intelligence technology, the rapid development of convergence neural network and face recognition technology has made great progress. This article mainly discusses the research on face recognition algorithms at home and abroad, and expounds the importance of face recognition in logistics. Analyze the feasibility and benefits of deep learning methods for face recognition in logistics. By comparing the pros and cons of feature recognition principles such as SVM, YOLO, SSD, CNN, etc., the improved CNN is finally selected as the key face recognition model in this article. The purpose of this article is to study the face recognition of the CNN model to improve the recognition accuracy and improve the detection rate, detection rate and accuracy.

REFERENCES

[1] Anna K. Bobak, Benjamin A. Parris, Nicola J. Gregory. Eye-movement strategies in developmental prosopagnosia and "super" face recognition[J]. Quarterly Journal of Experimental Psychology, 2017, 70(2):201-217.

[2] Rahim R , Af Riliansyah T , Winata H , et al. Research of Face Recognition with Fisher Linear Discriminant[J]. IOP Conference Series Materials Science and Engineering, 2018, 300(1):012037.

[3] Attwood A S , Easey K E , Dalili M N , et al. State anxiety and emotional face recognition in healthy volunteers[J]. Royal Society Open Science, 2017, 4(5):160855.

[4] Guo K , Wu S , Xu Y . Face recognition using both visible light image and near-infrared image and a deep network[J]. Caai Transactions on Intelligence Technology, 2017, 2(1):39-47.

[5] Min R , Kose N , Dugelay J . KinectFaceDB: A Kinect Database for Face Recognition[J]. IEEE Transactions on Systems Man & Cybernetics Systems, 2017, 44(11):1534-1548.

[6] Lu J , Liong V E , Wang G , et al. Joint Feature Learning for Face Recognition[J]. IEEE Transactions on Information Forensics & Security, 2017, 10(7):1371-1383.

[7] Peng C , Gao X , Wang N , et al. Graphical Representation for Heterogeneous Face Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(2):301-312.

[8] Best-Rowden L , Jain A K . Longitudinal Study of Automatic Face Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(1):148-162.

[9] Landi S M , Freiwald W A . Two areas for familiar face recognition in the primate brain[J]. Science, 2017, 357(6351):591.

[10] Grm K , ?truc, Vitomir, Artiges A , et al. Strengths and weaknesses of deep learning models for face recognition against image degradations[J]. Iet Biometrics, 2018, 7(1):81-89.

[11] Feng Q , Yuan C , Pan J S , et al. Superimposed Sparse Parameter Classifiers for Face Recognition[J]. IEEE Transactions on Cybernetics, 2017, 47(2):378-390.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Multiscale Global Channel Network for Edge Detection

1st Wang Zhang
School of Artifical Intelligence and Computer Science
Jiangnan University
Wuxi, China
forzhangwang @163.com

2nd Wei Fang
School of Artifical Intelligence and Computer Science
Jiangnan University
Wuxi, China
fangwei@jiangnan.edu.cn

3rd Jun Sun
School of Artifical Intelligence and Computer Science
Jiangnan University
Wuxi, China
junsun@jiangnan.edu.cn

4th Qidong Chen
School of Artifical Intelligence and Computer Science
Jiangnan University
Wuxi, China
cqd_mu@hotmail.com

***Abstract*—Neural network is used to fuse the spatial and channel information of each layer's local receptive fields to construct information features. But global long-range dependency is not effectively modeled, which leads to non-optimal discriminative feature representations. In this paper, we propose multi-scale global channel network (MSGC). We use self-attention mechanism to combine local features with their corresponding global dependencies, adaptively recalibrate the channel response, guide the network to ignore irrelevant information, and emphasize the correlation of relevant features. We evaluated the proposed method on BSDS500 dataset and NYUD dataset. MSGC achieves ODS F-measure of 0.815 on BSDS500, which is 0.9% higher than the existing technology.**

Keywords-Convolutional Neural Networks; Deep learning; Edge detection; Deep attention; Self-attention

I. INTRODUCTION

Edge detection aims to extract perceptually salient edges of natural images, which is important to high level computer vision tasks, such as image segmentation [1], [2], object detection/recognition[3], [4].

The early traditional methods include Sobel detector [5], widely used Canny detector [6],Structured Edges [7] and gPb [2]. CNN is used for edge detection, including DeepContour [8] and CSCNN [9]. HED [10] and RCF [11] supervise the predictions of different network layers. Richer convolution features are very effective for many visual tasks, but HED and RCF still do not explicitly use global context information, and do not directly impose constraints on adjacent pixel labels to enhance depth supervision. Therefore, we can improve the quality of network representation by explicitly modeling the dependency of channels.

Because convolution layer establishes the pixel relationship in the local neighborhood[12],[13], the modeling of long-range dependency is invalid. We add features of all positions to the features of each location.

II. RELATED WORK

This paper mainly studies edge detection and deep attention. We briefly review the related work in these two aspects.

A. Edge Detection

Edge detection is one of the most basic and challenging problems in computer vision.

These methods can be roughly divided into three categories: traditional edge operators, learning based methods and deep learning based methods. The traditional edge operator detects edges by detecting abrupt changes in intensity, color and texture. Sobel [5] applied thresholding the gradient of the image. The learning based method uses hand-crafted features. Arbeláez *et al.* [1] combined local clues into a global framework. In recent years, advanced results have been obtained by using deep learning to extract depth features automatically. Xie and Tu [10] proposed an end-to-end model for in-depth monitoring different scale features of side outputs. Liu *et al.* [11] connected the side outputs to all the convolution layers of VGG16 [14]. MSGC is based on RCF [11]. The above training strategy does not explicitly use context information and impose constraints on adjacent pixel labels. We use global features to enhance context modeling for multiscale side outputs.

B. Deep Attention

The attention mechanism aims to emphasize the important areas and filter irrelevant information. It has been successfully applied to various visual tasks, such as classification [15] and detection [16]. PSANET [17] adaptively linked each location in the feature map with other locations. Senet [13] and Genet [18] recalibrated channel dependencies by rescaling different channels. However, the feature fusion method is not effective enough. In this paper, addition fusion is used to model the global context more effectively.

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00033

III. METHODS

A. Overview

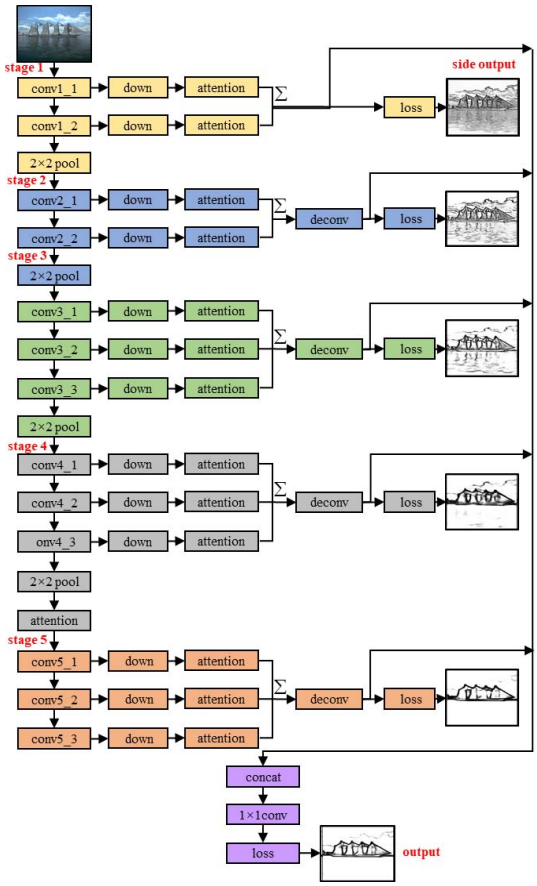


Figure 1. The architecture of MSGC.

VGG16 network is composed of 13 convolution layers, 3 fully connected layers and 5 downsampling layers. We make following changes to VGG16: (1) because the fifth pooling layer produces a too fuzzy prediction map to be used, the fifth pooling layer and all fully connected layers are discarded, (2) a downsampling layer is connected to each convolution layer to extract different scale features, (3) fuse the multiscale features of each stage, then the deconvolution layer up-sample the fused features, (4) a convolution layer is used to fuse the side outputs.

Let (X,Y) denotes one sample of input training data set T, where $X=\{x_i, i=1,\cdots,|X|\}$ is a raw input image and $\mathrm{Y}=\{\mathrm{y}_i,\mathrm{i}=1,\cdots,|\mathrm{X}|\}$, $\mathrm{y}_j\in\{0,1\}$ is the corresponding ground truth edge map. The training loss for every image is formulated as

$$l(X;W)=\alpha\sum_{i\in Y_+}\log P(y_i=1\mid X;W)+\beta\sum_{i\in Y_-}\log(1-P(y_i=0\mid X;W)), \quad (1)$$

where $\alpha=\lambda\cdot\frac{|\mathrm{Y}_-|}{|\mathrm{Y}_+|+|\mathrm{Y}_-|}$, $\beta=\frac{|\mathrm{Y}_+|}{|\mathrm{Y}_+|+|\mathrm{Y}_-|}$, Y_+ and Y_- denote the edge and non-edge ground truth label sets respectively, λ is to automatically balance the loss between positive/negative classes, and W denotes all the network layers parameters. The final loss can be obtained by further aggregating these generated edge maps, *i.e.*,

$$\mathrm{L}(\mathrm{W})=\sum_{j=1}^{5}l(X^j;W)+l(X^{fuse};W), \quad (2)$$

where X^{j} denotes the edge map of stage j and X^{fuse} denotes the edge map of fusion layer.

Traditional CNNs has a local receptive field, so the generated local features may cause potential differences between features of pixels with the same label. We study the self-attention mechanism of establishing association between features. First, the global context information is captured. Then, the learned global features are input into the channel self-attention module. The self-attention module helps to adaptively combine local features with their global context, and can gradually filter out noise by emphasizing useful information. The overview of the proposed architecture is depicted in Figure 1.

B. Global Channel Self-attention Modules (GCM)

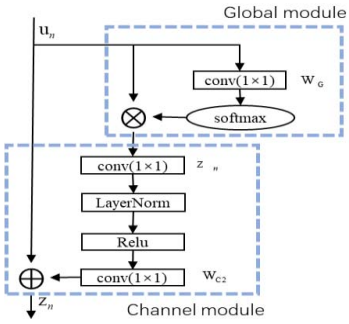


Figure 2. Architecture of the GCM.

Firstly, we use 1×1 convolution $\mathrm{W_G}$ and softmax function to obtain attention weights, and compute a global context attention map S. Then we recalibrate the channel response through 1×1 convolution $\mathrm{W_c}$. Finally, we aggregate the global context features to the features of each location by addition. We use $\mathrm{U}=\{\mathrm{u}_n,\mathrm{n}=1,\cdots,\mathrm{N}\}$ as an input feature map, where $\mathrm{N}=\mathrm{H}\times\mathrm{W}$ is the number of positions in the feature map. Our global attention map is formulated as follows:

$$\mathrm{S}=\sum_{n=1}^{N}\frac{\mathrm{f}(u_n)}{C(U)}u_n, \quad (3)$$

where n lists all possible locations, $\mathrm{f}(u_n)=\mathrm{e}^{W_g\mathrm{u}_n}$ is an embedded Gaussian function to calculate the similarity in the embedding space, $C(U)=\sum_{\mathrm{m}=1}^{N}\mathrm{e}^{W_g\mathrm{u}_m}$ is a normalization factor.

In order to make the GCM lightweight, we use the bottleneck transform module. We add layer normalization in bottleneck transformation before ReLU layer to simplify optimization, and also play the role of regulation, which is illustrated in Figure 2. We denote $Z=\{z_n, n=1,\cdots, N\}$ as the output feature maps of our attention module, the complete GCM can be expressed as:

$$z_n = u_n + W_{C2}\,\mathrm{Re}lu(LN(W_{C1}S)). \quad (4)$$

IV. EXPERIMENTS

A. Datasets

BSDS500 contains 200 training, 100 verification and 200 test images. We expanded the training set and verification set with rotation, flipping, scaling. We mix the enhanced data of with the flipped Pascal VOC context dataset [19] as training data with 49006 training.

The NYUD [20] dataset consists of 1449 pairs of aligned RGB and depth images. We only use the RGB part. We split the NYUD dataset into 381 training, 414 validation, and 654 test images[21], and expand them by randomly flipping, scaling and rotating.

B. Implementation Details

We implement MSGC using PyTorch. The VGG16 pretrained on ImageNet [22] is used to initialize MSGC. The threshold λ used for loss computation is set as 1.1 and 1.2 for BSDS500 and NYUD dataset, respectively.

SGD optimizer randomly extracts 10 images in each iteration, and the global learning rate is set to 1e-6, which decreases 10 times after every 10K iterations. Momentum and weight decay are set to 0.9 and 0.0002, respectively. We do a total of 40K iterations. All experiments in this paper are performed with NVIDIA 1080 GPU.

Before evaluation, we used non-maximum suppression (NMS) to refine the edge. The maximum allowed tolerance between edge predictions and ground truth for BSDS500 and NYUD dataset are set to 0.0075 and 0.011 respectively.

C. Comparison with Other Works

Performance on BSDS500： Our experimental results with several state-of-the-art edge detection networks on BSDS500 are summarized in 0 and Figure 3.

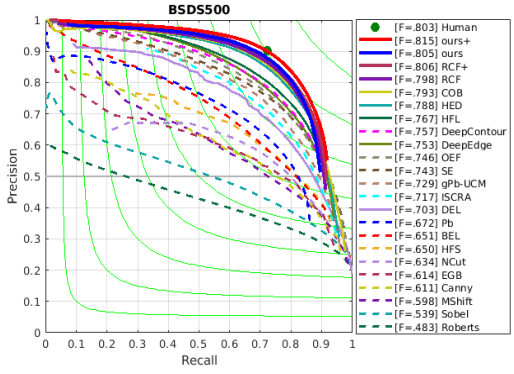


Figure 3. The p-r curves of MSGC and other works on BSDS500 dataset.

TABLE I. The comparison with other methods on BSDS500 dataset. +indicates trained with additional PASCAL VOC Context dataset.

Methods	ODS	OIS	AP
Human	0.803	0.803	-
Canny [6]	0.611	0.676	0.520
SE [7]	0.743	0.763	0.800
OEF [23]	0.746	0.770	0.820
DeepEdge [6]	0.753	0.769	0.784
DeepContour [27]	0.757	0.776	0.790
HFL [29]	0.767	0.788	0.795
HED [10]	0.788	0.808	0.840
CEDN+ [25]	0.788	0.804	-
RDS [29]	0.792	0.810	0.818
RCF [11]	0.798	0.815	-
RCF [11]+	0.806	0.824	0.840
DeepBoundary [27]	0.789	0.811	0.789
DeepBoundary+ [27]	0.809	0.827	0.861
MSGC	**0.805**	**0.822**	0.834
MSGC+	**0.815**	**0.834**	**0.866**

As shown in the results, GCM can actually improve the performance of edge detection. Figure 4. shows a comparison of edge maps from MSGC and RCF before NMS. MSGC can effectively eliminate most of the blurred and noisy boundaries and produce clearer image edges.

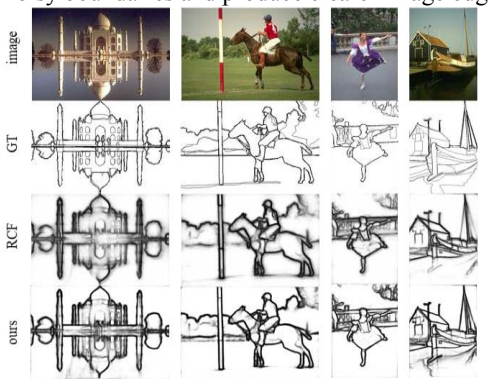


Figure 4. Comparison of edge maps before NMS on BSDS500 dataset.

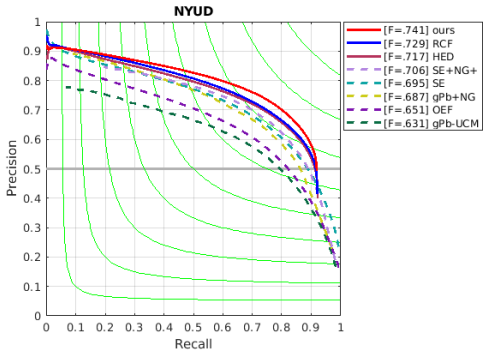


Figure 5. The p-r curves of MSGC and other works on NYUD dataset

TABLE II. Comparison with other methods on NYUD dataset.

Methods	ODS	OIS	AP
gPb-UCM [1]	0.631	0.661	0.562
gPb+NG [21]	0.687	0.716	0.629
OEF [23]	0.651	0.667	-
SE [7]	0.695	0.708	0.679
SE+NG+ [28]	0.706	0.734	0.738
HED [10]	0.717	0.732	0.734
RCF [11]	0.729	0.742	-
LPCB [29]	0.739	0.754	-
MSGC	**0.741**	**0.759**	**0.740**

Performance on NYUD： TABLE II. and Figure 5. show the quantitative results of MSGC compared with several recent methods. MSGC achieves the best performance of ODS F-score 0.741, which proves the effectiveness of MSGC.

V. CONCLUSION

In this paper, we introduce a deep attention architecture to complete the edge detection task. It combines different levels of global information with GCM to model long-range dependency effectively. Finally, a dynamic channel-feature recalibration is performed to filter the noisy regions and help the network focus on the relevant areas in the image. MSGC is compared with more than 10 edge detection methods on BSDS500 dataset and NYUD dataset, and MSGC provides accurate and reliable edge detection.

ACKNOWLEDGMENT

We wish to thank every member of the team for their efforts.

REFERENCES

[1] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. IEEE TPAMI,33(5):898–916, 2011.

[2] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interac-tive foreground extraction using iterated graph cuts. In IEEE Trans. Graph., volume 23, pages 309–314. ACM, 2004.

[3] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, pages 91–99, 2015.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, volume 1, pages 886–893. IEEE, 2005.

[5] J. Kittler. On the accuracy of the sobel edge detector. Image and Vision Computing, 1(1):37–42, 1983.

[6] J. Canny. A computational approach to edge detection. IEEE TPAMI, 8(6):679–698, 1986.

[7] P. Dollár and C. L. Zitnick. Fast edge detection using structured forests. IEEE TPAMI, 37(8):1558–1570, 2015.

[8] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang. DeepContour: A deep convolutional feature learned by positivesharing loss for contour detection. In IEEE CVPR, pages 3982–3991, 2015.

[9] J.-J.Hwang and T.-L.Liu. Pixel-wise deep learning for contour detection. arXiv preprint arXiv:1504.01989, 2015.

[10] S. Xie and Z. Tu. Holistically-nested edge detection. In IJCV. Springer, 2017.

[11] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Jia-Wang Bian, Le Zhang, Xiang Bai, Jinhui Tang. Richer Convolutional Features for Edge Detection. IEEE TPAMI, 2019.

[12] X.Wang,R.Girshick,A.Gupta,andK.He. Non-local neural networks. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018.

[13] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[15] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.

[16] H. Li, Y. Liu, W. Ouyang, and X. Wang, "Zoom out-and-in network with map attention decision for region proposal and object detection," International Journal of Computer Vision, vol. 127, no. 3, pp. 225–238, 2019.

[17] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, and J. Jia, "PSANet: Point-wise spatial attention network for scene parsing," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 267–283.

[18] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In Advances in Neural Information Processing Systems, pages 9423–9433, 2018.

[19] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Y uille. The role of context for object detection and semantic segmentation in the wild. In IEEE CVPR, pages 891–898, 2014.

[20] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In ECCV, 2012.

[21] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In CVPR, 2013.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei. Imagenet: Alarge-scalehierarchicalimagedatabase. In IEEE CVPR, pages 248–255. IEEE, 2009.

[23] S. Hallman and C. C. Fowlkes. Oriented edge forests for boundary detection. In IEEE CVPR, pages 1732–1740, 2015.

[24] G. Bertasius, J. Shi, and L. Torresani. High-for-low and low for-high: Efficient boundary detection from deep object features and its applications to high-level vision. In IEEE ICCV, pages 504–512, 2015.

[25] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang. Object contour detection with a fully convolutional encoder-decoder network. arXiv preprint arXiv:1603.04530, 2016.

[26] Y. Liu and M. S. Lew. Learning relaxed deep supervision for better edge detection. In IEEE CVPR, pages 231–240, 2016.

[27] Y. Wang, X. Zhao, and K. Huang. Deep crisp boundaries. In CVPR, 2017.

[28] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-dimages for object detection and segmentation. In ECCV, pages 345–360. Springer, 2014.

[29] R. Deng, C. Shen, S. Liu, H. Wang, and X. Liu. Learning to predict crisp boundaries. In ECCV, pages 562–578, 2018.

Classroom monitoring system based on facial expression recognition

Boxuan Zhang,Dandan Wei,Qianying Zhang,Wenyu Si,Xiang Li,Quanyin Zhu*
Faculty of Computer&Software Engineering,Huaiyin Institute of Technology,Huaian,China
*Corresponding author's e-mail:hyitzqy@qq.com

***Abstract*—Facial expressions are important information that reflects human emotions. Recognizing dynamic expressions of students in class, 8 kinds of emotions are selected for application: positive emotions: "happy"; negative emotions: "disgust , Sadness, doubts, contempt, anger"; neutral emotion: "focus, surprise". [1] In this design, the classroom performance scoring system in normal hours is split into four functions: wireless network list acquisition and verification, face recognition, emotion analysis, and scoring record storage. On this basis, SVM and Softmax are used. Facial expression recognition and a data storage database is designed to realize the function of an intelligent scoring system for classroom performance points. To solve this problem, an expression recognition method combining pyramid convolutional neural network and attention mechanism is proposed.**

Keywords-component; facial expression recognition; pyramid convolution; attention mechanism

I. INTRODUCTION

For the past few years, the continuous expansion of colleges and universities has resulted in a shortage of teaching resources to some extent, and the workload of schools is heavy, in the process of college education and student work management, the final exam is an important content, and class performance scores also account for a part of it. However, some problems have also arisen, such as incomplete recognition by teachers, the inability of teachers to grade students' class status during class, and the long time consuming scoring. This led to some mistakes in the scoring of teachers 'classroom performance. Therefore, a classroom monitoring system based on facial expression recognition and behavior prediction was developed to meet the actual needs of the current final assessment, such as schizophrenia and post-traumatic stress disorder.[2]

Therefore, we propose a facial expression recognition network based on the attention model to achieve the best detection effect by letting the system learn the important parts and ignore the unimportant parts. We will verify the superiority of this method in commonly used data sets.

This project is a classroom monitoring system based on facial expression recognition and behavior prediction.

II. RELATED WORK

A. Expression recognition from the classroom perspective

Facial expression,which is the most direct and effective emotion recognition mode, has many application scenarios. For instance, people use facial expression to detect the fatigue driving. However, we apply this technology to students' classrooms, and use this to help teachers make more accurate ending points. The basic expressions that people put forward at the beginning are:angry, scared, Disgust, happy, sad, surprised and neutral.

Traditional hand-designed features and even shallow learning features can no longer adapt well to various interference factors that are not related to expressions in the real world, such as light changes, different head postures, and facial blocking.

On this basis,the study of attention mechanism started to be applied to facial expression recognition.

For facial expression recognition in real scenes, direct occlusion caused by objects or indirect occlusion caused by factors such as illumination and posture changes is one of the inherent challenges of expression recognition. Because the occlusion in the real scene is complex and diverse, the feature reconstruction method relies on a large number of training data under different occlusion conditions to have better results, and the reconstruction of face details is not ideal. An expression recognition method that does not rely on the detection of key points on the face is needed to recognize expressions in real scenes.

B. Attention mechanism

An important feature of the human visual system is that it does not try to process the entire scene at once, but selectively focuses on the target area that needs attention to obtain the detailed information of the target, while suppressing other useless information, which greatly improves the visual information processing Rate.

Zhang et al. [3] proposed an attention layered bilinear pooling residual network for expression recognition. This method uses the channel attention mechanism to explicitly model the importance of each channel, and assigns different output feature maps. The weight of, locate the salient area according to the weight value. Li et al. [4] used the spatio-temporal attention mechanism to recognize facial micro-expression, and the temporal attention module was used to learn the motion information of the expression sequence, focusing on the more discriminative frames in the expression sequence.Wang et al. [5] proposed a new type of regional attention network, which extracts the features of each region through the backbone convolutional network, and weights the attention feature information to improve the accuracy of facial expression recognition under occlusion and posture changes. Amir et al. [6] proposed a deep attention center loss (Deep Attentive Center Loss, DACL) method. The proposed DACL integrates an attention mechanism and uses the intermediate spatial feature map extracted by CNN as the context to estimate and feature Importance-related attention weights to adaptively select a subset of important feature elements to enhance discrimination.

Based on the method proposed by Wang [5], this paper proposes a global attention module. The attention

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00034

mechanism allows more prominent features to be selected according to needs. The global attention module in this paper can better solve the problem of real scene expression recognition.

C. Pyramid Convolution

In the training process of deep neural networks, image feature information is extracted through convolution operations, and the spatial features that can be learned by convolution kernels of different sizes are not the same. For small targets and targets with noise, detailed feature information is very important, and pixel-level deviations often lead to errors in recognition. Pyramidal Convolution (PyConv)[6]can process input information through multiple convolution kernels of different scales. The main advantage of PyConv is multi-scale processing, with different spatial resolutions and depths. Compared with standard convolution, PyConv can expand the receptive field of the convolution kernel without increasing additional costs. Pyramid convolution can capture the diversity of expression features and the variability of their scales, and maintain the continuity of facial action units.

III. EASE OF USE

A. Sub-image generation

The neurons in each layer of the convolutional neural network are arranged in three dimensions: width, height, and depth. The width and height are well understood, because the convolution itself is a two-dimensional template. However, the depth in the convolutional neural network refers to the third dimension of the activated data volume, not the depth of the entire network. The depth of the entire network refers to the number of layers of the network.

We will see that the neurons in the layer will only be connected to a small area in the previous layer, instead of being fully connected.

B. Pyramid Convolutional Network

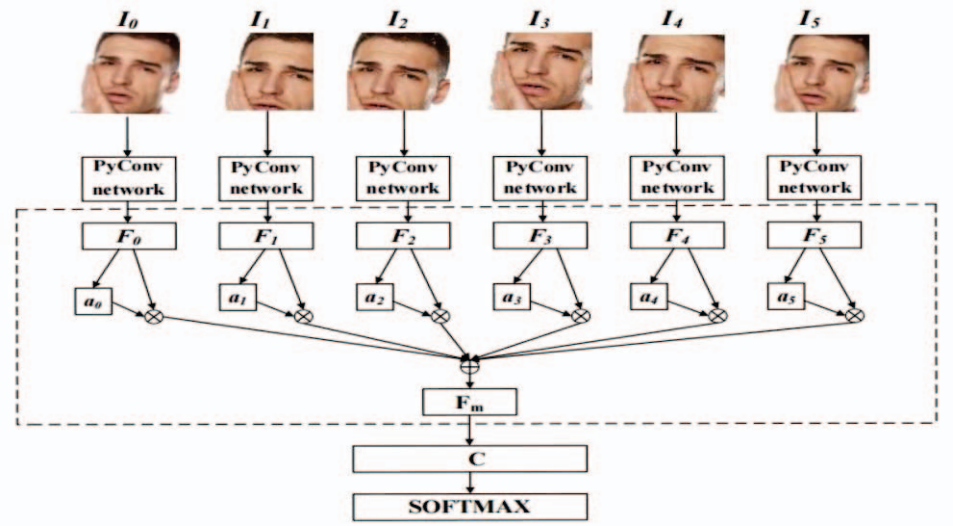


Fig.1 PyConv-Attention network mod

The key problem of facial expression recognition is to find the expression feature areas with prominent expression changes. The attention mechanism is widely regarded as a method to help solve this type of problem. Studies have shown that the mouth, eyes, eyebrows and nose are formed Different action units, the combination of these facial units forms face expressions. PyConv is a pyramid convolution unit. As shown in Figure 2, it is composed of convolution kernels of different sizes and different depths. As the size of the convolution kernel increases, its depth decreases accordingly. These convolution kernels can capture Different levels of detail features in the image. For the centi-word tower convolution unit, due to the use of convolution kernels of different depths, a grouped convolution method is adopted. The input feature map is split into several parts, and the convolution kernels of different depths are used for each group of input feature maps. Feature extraction. When the grouping is 1, it is a standard convolution, where the depth of the convolution kernel is equal to the number of channels of the input feature map.

C. Attention module

The proposed global attention module is shown in the dashed box in Figure 1. After the original image and sampled sub-images are extracted from the centripetal convolutional network, they are sent to the global attention module to calculate the image's value.

D. Attention module

The proposed global attention module is shown in the dashed box in Figure 3. After the original image and sampled sub-images are extracted from the centripetal convolutional network, they are sent to the global attention module to calculate the image's value through a fully connected layer and Sigmod activation function. The feature weight ai is finally weighted and summed to obtain a global feature representation Fm. We denote the original image as I_0, the sub-images as I_0, I_1, ..., I_k, and

the backbone network as $r(I^*;\theta)$. The feature set X of the image I* is defined as the formula (1):

$$X\,[F_0, F_1, ..., F_k] = [r(I_0;\theta), r(I_1;\theta), \cdot, r(I_k;\theta)] \quad (1)$$

It is sent to the global attention module to use the fully connected layer and the Sigmoid activation function to calculate the attention weight.

The meaning weight is expressed as formula (2):

$$a_i = f(F_i^T, q) \quad (2)$$

Since there is an important feature information area in an expression image, we set the attention loss function to limit the attention weight of the original image and the sub-image. The loss function requires that the attention weight from the sub-image, yi, should be greater than the weight of the original expression image. The attention loss function formula is shown in(3):

$$L = \max\{0, \partial - (\mu_{max} - \mu_0)\} \quad (3)$$

Among them is the hyperparameter, which is set to 0.03 in this paper, 0 is the attention weight of the original image, and max represents the maximum weight of all sub-images.

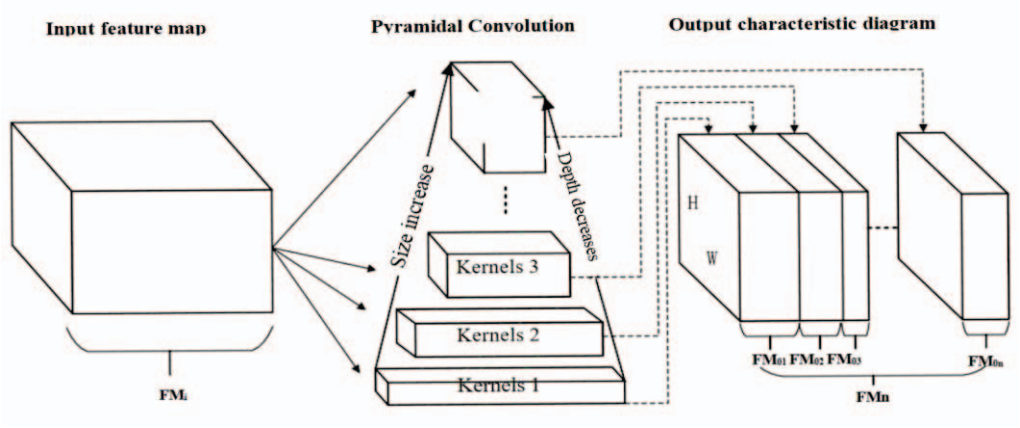


Fig.2 Pyramidal Convolution

IV. Experiments and Results

A. Pretreatment

Because the image sizes of different data sets are different, before training the model, the data needs to be preprocessed to adjust the size of all images to 224×224×3. Due to the differences in samples in different data sets, the convergence speed of the model on different data sets is different. In this experiment, the number of iterations on the CK+, RAF-DB, and AffectNet data sets are 100, 200, and 200, respectively. In the CK+ training process, the learning rate of the 40th and 80th rounds is attenuated with a decay rate of 0.9; in the training process of RAF-DB and AffectNet, the learning rate is attenuated with a decay rate of 0.9 every 50 rounds.

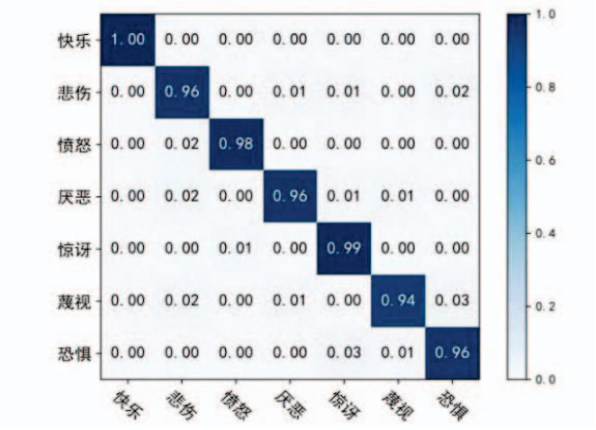


Fig.3 Confusion matrix of CK+

B. CK+ experiment

The confusion matrix of the proposed model on the CK+ data set is shown in Figure 5. For the six expressions of happiness, sadness, anger, disgust, surprise and fear, the recognition accuracy rate is 95%.

The expression "contempt" with the lowest recognition accuracy rate also reached 94%, and the recognition rate of the two expressions with obvious characteristics of "happy" and "surprise" reached more than 99%. The proposed model is compared with other mainstream methods on the CK+ data set, and the experimental results are shown in Table 1. It can be seen that the accuracy of the model proposed in this paper on the CK+ data set is 98.46%. Compared with the four methods Gabor [7], WLS-RF [8], PACNN [9] and SCAN [10], it is 6.27%, 4.16%, 1.43%, 1.15% higher, respectively.

TABLEI. Experimental Comparison Of CK+

Method	Accuracy(%)
Gabor	92.19
WLS-RF	94.3
pACNN	97.03
SCAN	97.31
PyConv-Attention Network	98.46

This method realizes facial expression recognition of feature maps with less overhead through channel dimension reduction and expansion convolution. The experimental results show that the method in this paper significantly improves the accuracy of facial expression recognition on the CK+ expression data set.This paper uses the residual network as the basic framework to design a facial expression recognition model that combines pyramid convolution and global attention.Pyramid convolution can learn multi-scale feature information and improve the nonlinear expression ability of the model; the attention mechanism can make the network pay more attention to important feature information and suppress noise interference. The proposed model has achieved 98.46% accuracy on the CK+ public expression data set.

REFERENCES

[1] AlNatour Ahlam,Gillespie Gordon Lee,Alzoubi Fatmeh. "We cannot stop smoking": Female university students' experiences and perceptions.[J]. Applied nursing research : ANR,2021,61:

[2] Yunxin Huang, Fei Chen, Shahe Lv, Xuedong Wang. Facial Expression Recognition: A Survey [J]. Symmetry, 2019,11(10).

[3] ZHANG A M, XU Y. Attention Hierarchical Bilinear Pooling Residual Network for Expression Recognition [J].Computer Engineering and Application, 2020, 56(23):161- 166.

[4] Gera D, Balasubramanian S. Landmark Guidance Independent Spatio-Channel Attention and Complementary Context Information based Facial Expression Recognition[J].Pattern Recognition Letters, 2021, 145:58-66.

[5] LI G H, YUAN Y F, Ben X Y, ZHANG J P. Spatiotemporal attention network for micro- expression recognition[J].Journal of Image and Gra-phics,2020,25(11):2380-2390.

[6] Duta I C, Liu L, Zhu F, et al.Pyramidal convolution: rethinking convolutional neural networks for visual recognition[EB/OL].(2020-6-20)[2021-5-9]

[7] Adil B, Nadjib K M, Yacine L. A novel approach for facial expression recognition[C]//2019 International Conference on Networking and Advanced Systems (ICNAS). IEEE, 2019: 1-5.

[8] Dapogny A,Bailly K,Dubuisson S. Confidence -weighted local expression predictions for occlusion handling in expression recognition and action unit detection[J]. International Journal of Computer Vision, 2018, 126(2): 255- 271.

[9] Li Y, Zeng J, Shan S, et al. Occlusion aware facial expression recognition using CNN with attention mechanism[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2439-2450.

[10] Gera D, Balasubramanian S. Landmark Guidance Independent Spatio-Channel Attention and Complementary Context Information based Facial Expression Recognition[J]. Pattern Recognition Letters, 2021, 145:58-66.

Research on Campus Physical Measurement System Based on CNN-LSTM

Mengzhong Wei,Yuting Fu,Chao Li
Huaiyin Institute of Technology
Computer&Software Engineering
Huai'an China
1455301587@qq.com

Yuanyuan Li,Wanli Feng,Quanyin Zhu*
Huaiyin Institute of Technology
Computer&Software Engineering
Huai'an China
*Corresponding Author's email:hyitzqy@qq.com

***Abstract*— In response to the 14th Five-Year Plan and the 2035 long-term goal outline, the goals of building smart communities and smart industries are clearly defined. Starting from the campus, there are many physical test projects in colleges and universities. Problem, so the campus physical test is transformable and innovative. This article first proposes the idea of building a smart body measurement system; then, analyzes the CNN-LSTM (Convolutional Neural Network-Long short term memor)algorithm that is mainly used in the construction of the system; secondly, it analyzes the accuracy of the recognition of specific action processes and the debugging process of the actual parameters. Carry out optimization and improvement; finally, make a report on the enforceability of the campus physical test system.**

Keywords CNN-LSTM; human pose estimation; smart application; Opencv

INTRODUCTION

Since the 14th Five-Year Plan and the 2035 long-term goal outline [1] was released, research on the construction of smart campuses and smart campuses have been increasing. In the traditional sense, smart campus projects such as college information file management platform, subject competition organization management platform, colleges and universities. The construction of public platforms, etc. focuses on the processing of campus information data in the context of colleges and universities. With the rapid development of the Internet of Things and artificial intelligence, under the initiative of schools and enterprises to combine production and education, many smart classrooms, smart restaurants, etc., integrate smart campus teaching equipment or The terminal operates in colleges and universities, and has obvious effects in assisting campus teaching and improving the quality of campus life. The construction of the campus physical measurement system proposed in this paper focuses on the construction of the human skeleton feature model and the accuracy of human action recognition. Analyze the mode of the action through the student's physical test video, refer to the actual physical test score, record the student's performance data, and finally get the physical test result.

RELATED WORK

A. CNN-LSTM algorithm basis

Feedforward Neural Network, also known as Multilayer Perceptrons, is the most classic and basic mode in deep learning [2]. The purpose of this algorithm is to find an optimal approximation function f, so as to obtain a mapping from input x to output y: y = f(x; θ), where θ represents the parameters of the neural network. Figure 1 is a schematic diagram of a typical one-layer fully connected forward neural network:

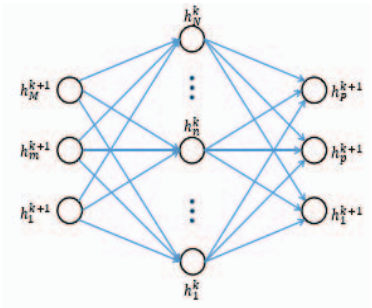


Figure 1. Model of neural network.

In the forward neural network, the information flow starts from the input X, passes through the intermediate f calculation, and outputs y. There is no feedback connection in this process, and there is no output of a certain layer as the input of the next layer. The forward neural network generally has a multi-layer structure. Assuming that l represents the number of neural network layers, and k represents the current level, the calculation mode of the forward neural network is:

$$h^k = \sigma(b^k + w^k h^{k-1}) \quad (1)$$

Among them, h^k represents the output of the Kth hidden layer, b^k represents the bias of the kth layer, w^k represents the corresponding weight, and σ represents the activation function, which is generally a sigmoid function or a tanh function.

When the parameters of the forward network θ are fixed, we can get the output of the neural network through Algorithm(1). The problem now is how to solve the parameters of the forward network. Like ordinary machine learning problems, we can solve the optimal parameter θ by minimizing the cost function J(θ). Among them, L represents the cost function of each sample, and f(x, θ) is the output of the neural network when the input is x.

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00035

Algorithm 2Multi-layer neural network forward propagation
Input: network depth l
Input: W(i),i ∈ {1, ..., l}, the weight of each layer of the model
Input: b(i),i ∈ {1, ..., l}, the bias of each layer of the model
Input: x, the input of the neural network
Output: ˆy, the output of the neural network
Output: J, the value of the loss function of the neural network

B. Convolutional Neural Network

Convolutional Neural Network (CNN) is a type of the previous neural network. LeCun et al. proposed that it has become the most commonly used and important in deep learning model after Hinton et al. won the first place in the ImageNet competition with CNN in 2012[3].The convolutional layer is the basis of the convolutional neural network, and the convolution operation is equivalent to using a sliding window to do a weighted average on an image. If our input data is x, the convolution kernel is f, and the output is y, here

$$\mathbf{x} \in \mathbb{R}^{H \times W \times D}, \mathbf{f} \in \mathbb{R}^{H' \times W' \times D \times D''}, \mathbf{y} \in \mathbb{R}^{H'' \times W'' \times D''} \quad (2)$$

Then the convolution operation can be expressed as:

$$y_{i''j''d''} = b_{d''} + \sum_{i'}^{H'} \sum_{j'}^{W'} \sum_{d'}^{D} f_{i'j'd} \times x_{i''+i'-1,j''+j'-1,d'd''} \quad (3)$$

The role of the Pooling layer is mainly to reduce the dimensionality of the output of the convolutional layer while retaining effective features. Pooling is mainly divided into two types: Max-pooling and Sum-pooling. Max-pooling operation can be expressed as:

$$y_{i''j''d} = \max_{1 \le i' \le H', 1 \le j' \le W'} x_{i''+i'-1,j''+j'-1,d}. \quad (4)$$

Sum-pooling and Max-pooling are similar to Sum-pooling, except that Max-pooling takes the maximum response. Local Response Normalization (LRN) normalizes the different channels at each position of the input data. The specific formula is as follows:

$$y_{ijk} = x_{ijk} \left(\kappa + \alpha \sum_{t \in G(k)} x_{ijk}^2 \right)^{-\beta}, \quad (5)$$

Among them, for each channel k, G(k) ⊂ {1, 2, ..., D} corresponds to the input subset. For LRN, input data and output data have the same dimensions.

For most machine learning problems, we can usually mix multiple models to get better generalization errors, but for deep neural networks, training multiple neural networks and then mixing the models often costs too much calculation resource. Dropout can be regarded as an approximation of model mixing. The method of Dropout is to set the input neurons to 0 with a certain probability p, so these neurons will not be counted in the forward and reverse processes.

C. Action recognition based on multi-resolution coding

The framework of the action recognition algorithm is shown in Figure 3-1. The algorithm is mainly composed of three parts: feature extraction, feature coding and action classification. Our action recognition framework combines traditional features and deep neural network features through different The coding method encodes them separately, and finally performs feature fusion and action classification [4].

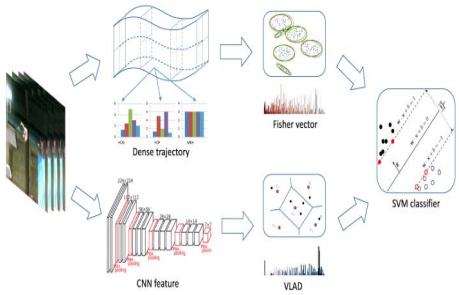


Figure 2. Framework diagram of multi-resolution coding action recognition.

LCD (Latent Concept Descriptor) is a feature extraction algorithm that encodes the features of neural networks through the traditional bag-of-words model [5]. Generally speaking, when we use CNN for feature extraction, we will use the features of the last fully connected layer as the input features. However, LCD encodes the features of the last layer of convolution, because this layer retains the input spatial information.

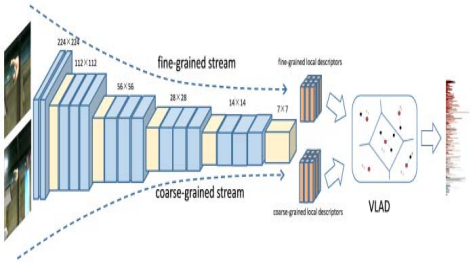


Figure 3. Multi-resolution LCD algorithm

First, we combined LCD features with Two-stream CNN. The original LCD feature only encodes one of the two-stream CNNs of Two-stream CNN, that is, a single frame image is used as the input of the neural network to extract high-level semantic information. Therefore, the original LCD can only obtain static features such as scenes in the video. In order to be able to extract the dynamic features of the video, we did the same operation on the optical flow neural network, which is to treat the

pool5 layer as a local feature and then encode it. The two CNNs in Two-stream CNN will be respectively encoded with LCD, and the extracted features will be feature fusion at the end.

In this section, we will introduce how to detect a significant area in a frame of image [6]. In identifying problems, usually, only one person is doing a certain action in a video. So, generally speaking, The most prominent area in the image includes the most important part of the frame. Our algorithm independently extracts features for this part during LCD encoding, which is equivalent to increasing the weight of the salient area in the final video feature. Given a whole input image I, we first use the EdgeBox algorithm to detect its candidate area (Object Pro-posal)[7]. The detection of object candidate regions generally refers to an algorithm for quickly screening out regions that may contain objects during object detection. After all the candidate regions are obtained, the formula for determining the final saliency region is as follows:

$$b = \frac{1}{T}\sum_{t}^{T} \Pi(w_t b_t < \alpha S)\Pi(w_t b_t > \beta S) b_t \quad (6)$$

Among them, S represents the area of the input image, and α and β are manually determined parameters. In our experiment, we set their values to 0.1 and 0.6, respectively. This formula is equivalent to screening all candidate regions according to their area, filtering out the too large and too small regions, and averaging all the remaining regions, so as to get the most significant region in the end. Of course, we can also calculate the saliency map of the input image from this[8]:

$$saliency(Pxy) = \sum_{t}^{T} \Pi(p_{xy} \in b_t) S_{xy} \quad (7)$$

Among them, (x, y) represents the position of a pixel. The algorithm for the detection of the entire salient area is shown in Figure 4, which shows the original image in turn, with the 10 candidate areas with the highest score, the detection result of the salient area, and the visible image of the image. Of course, we only need to use the saliency region in our motion detection algorithm.

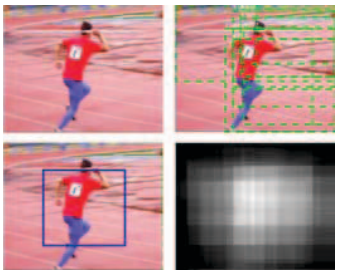


Figure 4. Significant area detection

EASE OF USE

A. Experimental results

In this section, we will first briefly introduce the data set used in the experiment: Olympic Sports. The Olympic Sports data set includes sports videos of various athletes collected from YouTube. The data set includes a total of 16 action categories (including weightlifting, bowling, basketball layup, diving, high jump, etc.), including a total of 783 videos . We use 649 videos as the training set, and the remaining 134 videos as the test set. Similarly, on this data set, the evaluation index is mAP.

B. Experimental results

In the data set we use Map[9] (Mean Average Precision) as the evaluation index. For ordinary classification problems, generally we will directly use accuracy as an evaluation indicator. AP (Average Precision) can be calculated with the following formula:

$$AP(\varepsilon) = \frac{1}{N_i}\sum_{k=1}^{N_i} P(k) \quad (8)$$

Among them, P(k) represents the accuracy (Preci-sion) when k videos are retrieved from the database by setting the threshold, and mAP can be expressed as the averagevalue of all APs, namely:

$$mAP = \frac{1}{M}\sum_{i=1}^{M} AP(\varepsilon) \quad (9)$$

C. Visualization of results

Filter out the fixed actions from Olympic Sports, use Openpose and Opencv tools to instantiate the above-mentioned action pattern recognition process, and get the action posture recognition diagram:

Figure 5. Action recognition diagram

D. Visualization of results[10]

Starting from the 14th Five-Year Plan proposed by the country, this article starts from the direction of smart applications and uses related ML algorithms to process video streams. The more basic ones are the popularization of CNN-LSTM algorithm forward algorithm and the detection of action pattern recognition algorithms. , Gave a detailed overview of the action recognition process and

saliency area detection in multi-resolution video analysis, and finally conducted an experiment on the collected instance Olympic Sports data set, and the class roughly inferred the action mode and movement correlation. This research has good application prospects in campus physical testing, which can save a lot of manpower and material resources and achieve intelligent physical testing. But it is worth noting that in practical applications, user identification and anti-cheating measures need to be improved.

IV. SUMMARY

Starting from the 14th Five-Year Plan proposed by the country, this article starts from the direction of smart applications and uses related ML algorithms to process video streams. The more basic ones are the popularization of CNN-LSTM algorithm forward algorithm and the detection of action pattern recognition algorithms. , Gave a detailed overview of the action recognition process and saliency area detection in multi-resolution video analysis, and finally conducted an experiment on the collected instance Olympic Sports data set, and the class roughly inferred the action mode and movement correlation. This research has good application prospects in campus physical testing, which can save a lot of manpower and material resources and achieve intelligent physical testing. But it is worth noting that in practical applications, user identification and anti-cheating measures need to be improved.

REFERENCES

[1]AlNatour Ahlam,Gillespie Gordon Lee,Alzoubi Fatmeh. "We cannot stop smoking": Female university students' experiences and perceptions.[J]. Applied nursing research : ANR,2021,61:

[2] Yunxin Huang, Fei Chen, Shahe Lv, Xuedong Wang. Facial Expression Recognition: A Survey [J]. Symmetry, 2019,11(10).

[3] ZHANG A M, XU Y. Attention Hierarchical Bilinear Pooling Residual Network for Expression Recognition [J].Computer Engineering and Application, 2020, 56(23):161- 166.

[4]Gera D, Balasubramanian S. Landmark Guidance Independent Spatio-Channel Attention and Complementary Context Information based Facial Expression Recognition[J].Pattern Recognition Letters, 2021, 145:58-66.

[5]LI G H, YUAN Y F, Ben X Y, ZHANG J P. Spatiotemporal attention network for micro- expression recognition[J].Journal of Image and Gra-phics,2020,25(11):2380-2390.

[6]Duta I C, Liu L, Zhu F, et al.Pyramidal convolution: rethinking convolutional neural networks for visual recognition[EB/OL].(2020-6-20)[2021-5-9]

[7] Adil B, Nadjib K M, Yacine L. A novel approach for facial expression recognition[C]//2019 International Conference on Networking and Advanced Systems (ICNAS). IEEE, 2019: 1-5.

[8] Dapogny A,Bailly K,Dubuisson S. Confidence -weighted local expression predictions for occlusion handling in expression recognition and action unit detection[J]. International Journal of Computer Vision, 2018, 126(2): 255- 271.

[9] Li Y, Zeng J, Shan S, et al. Occlusion aware facial expression recognition using CNN with attention mechanism[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2439-2450.

[10] Gera D, Balasubramanian S. Landmark Guidance Independent Spatio-Channel Attention and Complementary Context Information based Facial Expression Recognition[J]. Pattern Recognition Letters, 2021, 145:58-66.

Campus safety monitoring system based on deep learning

Qianying Zhang ,Jin Hang,Boxuan Zhang,Mengzhong Wei
Faculty of Computer & Software Engineering , Huaiyin Institute of Technology ,Huai'an,China
1532202289@qq.com

Quanyin Zhu*,Xiang Li
Faculty of Computer & Software Engineering , Huaiyin Institute of Technology ,Huai'an,China
*Corresponding author's email:hyitzqy@126.com

Abstract— **In view of the safety hazards existing on campus, a campus safety detection system based on deep learning behavior recognition, license plate recognition and speed detection is proposed and designed. The design divides the campus safety monitoring system into three functional modules: hazard behavior identification, license plate tracking and vehicle speed measurement, and on this basis, designs the database of storage information and the warning system of speed too fast to realize the monitoring function of campus safety.**

Keywords Deep learning; Campus safety; Behavior recognition

INTRODUCTION

In recent years, with the national attention to campus safety, campus security management has been increasing. At present, the safety and security work of our campus mainly depends on security personnel, although play a certain effect, but there are also many problems, such as in the course of campus safety management, there are security personnel can not find the students' dangerous behavior in the first time, can not identify the speed. There is a security risk. Therefore, the traditional supervision of campus safety through manpower is not enough. Therefore, it is necessary to develop a safety monitoring system for the identification of dangerous behavior and speed detection on campus.

This project is a campus safety monitoring system that integrates hazardous behavior identification, license plate tracking, and vehicle speed measurement.

RELATED WORK

A. Vehicle speed detection algorithm based on virtual coil

The vehicle speed detection [1] algorithm based on virtual coil is mainly divided into several processes. (1) Read traffic video: read the video using MATLAB platform [2].

(2) Set virtual coil: use the manually set virtual coil. Because this coil is flexible, it can be easily displayed on the first frame image.

(3) Calculate the gray difference between virtual coils VL1 and VL2: calculate the gray values of initial virtual coils VL1 and VL2, which are L1 and L2 respectively; Then calculate the gray values of virtual coils vl3 and vl4 in the next frame, which are L3 and L4 respectively. Thus, the VL gray difference m and N between frames can be determined, that is:

$$M = L3 - L1,\quad N = L4 - L2.$$

Then, two sets of values can be obtained by using the cyclic frame difference, which are stored in an array, and the absolute value of the difference is used for statistics. Therefore, the method of difference between each subsequent frame and the first frame can be used. (4)Virtual coil control: read in each traffic video, draw the two groups of values and the ordinal number of the video frame into a two-dimensional image and analyze it. According to the jump change of the gray difference, get the corresponding gray threshold, calculate the time length of the vehicle passing through the virtual coil, and finally calculate the speed of the vehicle.

B. License plate detection

Efficient Det is used as the vehicle detection network[3], because the detection speed and accuracy of the network are well performed on the ImageNet data set.

Considering the different imaging angles of license plate in real environment, a distortion corrected license plate detection network is proposed. The network can detect different degrees of distorted license plates in natural scenes, and correct the distorted license plates to a rectangular shape similar to the front view. In order to locate the license plate more accurately, a network based on residual block is designed to extract the license plate features, and the convolution layer in the network is set to 3x3. In order to prevent the license plate feature information from disappearing after pooling, only four steps of 2 are used × 2, which reduces the input dimension by 16 times. Finally, as shown in the figure, the detection module has two parallel convolution layers: (1) one is used to embed the probability, and the function is activated as softmax [4] (2) The other is used to regress affine parameters. In order to extract the features on the deformed license plate, first consider a hypothetical square whose size is fixed at the center of the license plate (x, y). If the target probability of the unit is greater than the given detection threshold, the affine matrix is established by using the regression parameters to convert the virtual square into the license plate area[5]. Therefore, the license plate can be easily extended to horizontally and vertically aligned objects. Use Mi = [Xi, Yi] t (where I = 1, 2, 3, 4) to represent the four corners of the annotated license plate, which are marked clockwise from the upper left corner; n1=[-0.5， 0.5]T， n2=[0.5， -0.5]T， n3=[0.5, 0.5]T， n4=[-0.5，0.5]T represent the corresponding vertex of the standard unit centered on the origin. The height h, width W

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00036

and network step of the input image are given through Ns = 24 (four maximum pool layers), and the network output characteristic diagram is given by P × Q It consists of 8 volumes, where p = H / N and q = w / NS. For each point unit in the characteristic graph, there are 8 values to be estimated: the first 2 values (V1-V2) are the probability of object / non object, and the last 6 values (v3-v8) are used to establish the local affine transformation TXY, which is obtained by equation (1):

$$T_{xy}(n) = \begin{bmatrix} \max(v_3, 0) & v_4 \\ v_5 & \max(v_6, 0) \end{bmatrix} n + \begin{bmatrix} v_4 \\ v_5 \end{bmatrix} \quad (1)$$

Txy(n) uses the learned parameters, the bounding box representing the predicted license plate position is affine transformed. The maximum functions V3 and V6 are used to ensure that the diagonal is positive (avoiding unnecessary mirroring or excessive rotation). In order to match the output resolution of the network, the MX points are rescaled according to the reciprocal of the network step, and re centered according to each (x, y) in the feature map. By using the standardized function (2):

$$A_{xy} = \frac{1}{\beta}\left(\frac{1}{N_s}\mathrm{m} - \begin{bmatrix} x \\ y \end{bmatrix}\right) \quad (2)$$

Among them, β Is a scaling constant that represents the edges of a virtual square. Set up β= 8.15 is the mean point between the maximum and minimum dimensions of the candidate license plate in the augmented training data segmented by the network stride [6]. In order to maintain a good compromise between accuracy and processing time, the maximum dimension 508 and the minimum dimension 128 are selected in this paper. Assuming that there is an object (license plate) at point (x, y), the first part of the loss function considers the error between the regular square of a distorted version and the standardized marking points of the license plate, which is given by equation (3):

$$f_{affine}(x, y) = \sum_{k=1}^{4} ||T_{xy}(n_k) \quad A_{xy}(m_k)|| \quad (3)$$

The license plates with tilt or even distortion is corrected to the front license plate by the distortion correction network, and the output license plate is used as the input of license plate recognition. In this way, the corrected license plate makes license plate recognition extremely easy, without considering the external influence such as angle, distortion or illumination, which improves the speed and accuracy of license plate recognition.

C. Extraction and filtering of behavioral feature

The first prerequisite for using the campus monitoring system to identify dangerous behavior of campus personnel is to be able to extract the behavior characteristics of the tested personnel, and then send the extracted behavior characteristics into the trained model to classify, obtain the basic classification results, and determine whether it is dangerous behavior. However, influenced by the dynamic change, lighting change, background clutter, time change and so on, it is the first problem to be solved to extract the behavior characteristics of the person being tested from the complex background. Unlike analytical images, video analytics have a larger amount of data and more complex conditions to analyze. Feature extraction can reduce the data dimension at the same time, analyze effective information, filter invalid information. There are several kinds of behavioral analysis feature extraction algorithms: shape feature, motion feature, space-time feature and mixed feature. The problem of features is mainly the features of time domain and frequency domain, which are extracted by sliding window. The sliding window contains two variables：size（size of window）、step（slip step）。Index times of window size 2 are generally selected in practical applications，$size = 2^{ceil(log2(2*f))}$, where f represents the sampling frequency of the sensor, and ai represents the combined acceleration. The calculation method is given by formula (1).

$$a_i = \sqrt{(a_i^x)^2 + (a_i^y)^2 + (a_i^z)^2} \quad (1)$$

In the time domain characteristics, $mean = \frac{1}{n}\sum_{n=1}^{n}(a_i - mean)^2$ represents the mean，$std = \sqrt{\frac{1}{n-1}}$ represents standard deviation，mod represents the mode，$max = \max(a_i)$ represents the majority，$min = \min(a_i)$ represents a minimum, $range = |max - min|$ represents the rang，$above\_mean = \sum_{n=1}^{n}\prod(a_i > mean)$ represents the mean point number of data, represents correlation coefficient，given by formula (2)：

$$\rho_{x,\ y} = \frac{\cos(x, y)}{\sigma_x \sigma_y} \quad (2)$$

SMA Represents the signal amplitude area, given by formula (3)：

$$SMA = \frac{\int_0^t |x(t)|dt + \int_0^t |y(t)|dt + \int_0^t |z(t)|dt}{t} \quad (3)$$

Because the scene of behavioral feature extraction often contains crowded people, resulting in a large number of noise that affects the recognition, therefore, a feature filtering method based on cover[7]is adopted. The construction of the mask is shown in Figure 1:

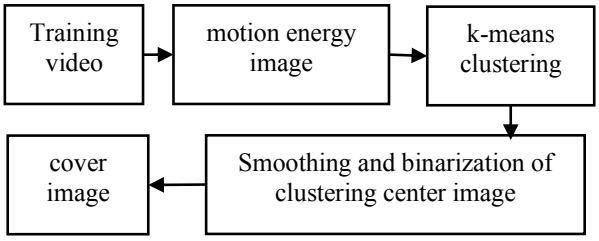


Figure 1. Construction process of shield

D. Classification of mask behavior

After feature filtering, behavior recognition becomes easier. The process of behavior recognition is as follows:
(1) The dangerous behavior images are input and the data are marked in the network model training. The location

behavior information of the detected personnel predicted by the model is used as the input of the recognition model to train the recognition model.

(2) The minimum circumscribed rectangle of the image to be judged is expanded in the left and right directions. The open window time length is T (representing the continuous time of the behavior). The video motion image MEI, motion history image MHI and optical flow energy image FEI are calculated in the window.

(3) The matching mask image is found from the behavioral mask of the training results, and the mask is placed on the detection window, MHI and FEI are retained, and the non-coverage area is removed to realize feature filtering.

(4) The behavioral characteristics are classified by the method based on trainers, so as to judge the behavioral characteristics.

Feature classification has two methods: direct classification and time domain state space fusion model. Dynamic Time Planning Adjustment (DTW) [8]is similar to the randomness of speech signal. The action behavior of different people in different scenes in video is different, and there is no exactly the same time length. Therefore, the time axis should be distorted and bent unevenly when matching with the established recognition model. Align video features with template features. DTW is essentially a time warping function that satisfies certain conditions to describe the time correspondence between the input model and the training model, so as to find the minimum cumulative corresponding warping function when two templates match.

REFERENCES

[1]Seda Kul, Isabek Tashiev, Ali Sentas, Ahmet Sayar:Event-Based Microservices With Apache Kafka Streams: A Real-Time Vehicle Detection System Based on Type, Color, and Speed Attributes. IEEE Access 9: 83137-83148 (2021)

[2]Liyan Zhao, Aize She:Design of Virtual Experiment Platform Based on MATLAB Simulation Technology. DPTA 2019: 335-341

[3]Riadh Ayachi, Mouna Afif, Yahia Said, Abdessalem Ben Abdelali:Drivers Fatigue Detection Using EfficientDet In Advanced Driver Assistance Systems. SSD 2021: 738-742

[4]Zhang Fan, Wang Xiaodong, Hao Xianpeng.Intelligent vehicle character recognition based on edge features [J].Automation and instrumentation, 2020 (6) : 11-4,20.

[5]Hae-Chan Noh, Jae-Pil Heo:Mutually Orthogonal Softmax Axes for Cross-Domain Retrieval. IEEE Access 8: 56491-56500 (2020)

[6] Caicai Li, Hao Jin, Huihui Bai:Ratio Test for Time-Varying Variance of Mean Change Point Based on Big Data Analysis. CSIA (1) 2021: 369-376

[7]Guo Ping.Video-based human behavior analysis.2012 Beijing Jiaotong University, PhD paper.

[8]Li Dongdong, Zhang Limin, Deng Xiangyang, Jiang Jie. Dynamic gesture recognition based on multi-feature fusion [J]. Computer applications and software, 2021,38 (08) : 214-219.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Parallelization implementation of topographic viewpoint filtering algorithm based on terrain viewshed using MPI and OpenMP

Yiwen Wang, Tao He, Wanfeng Dou
School of Computer & Elelctronic Information, Nanjing Normal University, Nanjing 210023
Nanjing 210023, Jiangsu, China
douwanfeng@njnu.edu.com

***Abstract*—The problem of multi-point visibility is an important part in terrain visibility analysis. It is widely used in military, urban planning, protection of endangered animal and other fields. Siting observer point is a kind of multi-point viewshed problem. It is generally abstracted as selecting the least number of viewpoints on a given terrain to maximize the joint viewshed covered by them. A candidate viewpoint filtering method proposed by Wang et al. to effectively solve this problem, includes k-means clustering of candidate viewpoints, calculation of contribution degree of each viewpoint, and ranking of viewpoints on contribution in each cluster. But this method is very time-consuming. Therefore, this paper adopts the MPI parallel program framework to implement the parallelization for the k-means algorithm, and uses OpenMP to realize the parallelization of the ranking process of view contribution of candidate viewpoints and the calculation process of view contribution of each viewpoint. The experimental results show that our parallelization scheme of viewpoint filtering can greatly reduce the calculation time and improve the efficiency of the algorithm.**

Keywords- multi-point viewshed analysis; observer siting; viewpoint filtering; parallelization

I. INTRODUCTION

Terrain visibility analysis uses computer graphics technology and computer geometry principle to solve the problem of visibility between target point set and observation point set on terrain. It is an important part of geospatial analysis [1]. Visibility issues can generally be divided into two categories: The first category is the calculation of visibility information of a given monitoring point. This information mainly includes the visual size and range of monitoring points of different properties and types[2]. The second category is to use these visual information to solve different application problems, such as location planning, path planning, real-time roaming and others[3].

Visibility analysis has been widely used in the process of environmental resource management, urban planning and military activity analysis, such as optimal path planning[4] , forest fire monitoring[5], communication tower positioning[6], military observation points siting[14],etc. These applications can be attributed to multi-point viewshed analysis. However, unlike single-point viewshed analysis that only needs to calculate the viewshed of a single point, multi-point viewshed analysis needs to obtain the largest or smallest joint viewshed by calculating the viewshed of a set of points. Therefore, the problem of observation point siting based on multi-point viewshed analysis is a combinatorial optimization problem, because it involves the optimal combination of multiple viewpoints and multiple target points, so it is also an NP problem[7]. The complexity of solving such a problem will increase exponentially with the increase of the number of candidate viewpoints.

At present, the viewpoint filtering method can effectively solve the multi-viewpoint viewshed. It first obtains some feature points on the terrain as candidate viewpoints, then evaluates these candidate points, and filters out some viewpoints with low contribution or poor quality until the required number of viewpoints is met[12-13]. In the problem of siting observation points, it is necessary to calculate the viewshed of a large number of terrain points, which is very time-consuming. Even if the terrain feature points are selected as the candidate viewpoints, the number is still very large. Therefore, it is of great significance to study the parallelization method and its solution of multi-viewpoint viewshed analysis. However, the current research mainly focuses on the parallelization of single-point viewshed computation, such as the parallelization of X-draw algorithm[14]. In the fast filtering method of terrain point, before performing the filtering operation, all the candidate points should be clustered by k-means method[15] , and then the terrain points in each cluster are sorted according to their viewshed quality. Although the feature points extracted from the whole terrain are used as candidate points, the number of these candidate observation points is still very large. Therefore, clustering operation and visual quality calculation of each terrain point takes a lot of time. This paper considers the parallelization of each step in the filtering method to improve the computational efficiency.

II. TERRAIN POINT FILTERING METHOD AND ITS FRAMEWORK OF PARALLEL METHOD

A. Terrain point filtering method

The main idea of the fast filtering method for candidate viewpoints is given as follows [12-13]. The candidate viewpoints are clustered by K-means algorithm based on geometric distance, and than N clusters are obtaind. Then, the viewpoints in each cluster are sorted according to their viewshed quality. Then the viewpoints with the lowest viewing quality in all clusters are filtered cyclically until the number of viewpoints in all cluster meet the predetermined number requirements. The fast filtering method for candidate viewpoints includes the following steps:

1)Topographic feature point extraction. With the acquisition of high precision terrain data, the number of terrain grid points is also increasing. If every point on the entire terrain is treated as a candidate viewpoint, it will inevitably lead to a large amount of time overhead. Therefore, using topographic feature points as candidate viewpoints can greatly reduce the time cost of location optimization calculation. Available topographic features

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00037

include mountaintops, ridge points, saddle points, flat locations[16],etc. The research shows that the topographic feature points have higher visibility than the general grid points in the terrain.

2)Clustering all topographic feature points. According to the specific requirements of the actual number of site selection plans, we can determine the total number N of clusters. In this paper, K-means clustering algorithm is used to cluster the terrain feature points, and the clustering centers of multiple clusters and each cluster are obtained.

3)Sorting the terrain feature points by visual quality in each cluster. The evaluation index is calculated according to the viewshed quality of the viewpoints in each cluster and sorted according to the value. Here, all viewpoints in clusters are sorted in descending order of their evaluation index.

4)Filter the feature points with the lowest quality of viewshed. Comparison is made on the viewshed quality of feature points in all clusters. If a feature point in a cluster is with the lowest viewshed quality in all clusters, this point will be deleted. Otherwise, it will be inserted into another cluster that is not with the smallest view quality.

5)Repeat steps 3-4 until the numbers of points in each cluster meets the expected requirements.

The fast filtering method of candidate feature points traverses all clusters until the number of viewpoints in the cluster satisfies the termination condition. Please pay attention to that when repeating Step 3, just reassess the viewshed quality of all feature points in the cluster where new feature points are inserted, rather than all clusters.

B. Parallel framework of topographic point filtering algorithm

The terrain point filtering algorithm is divided into three parts: the clustering of terrain points, the sorting of terrain points in the cluster, and the filtering of terrain points.

Due to the large number of the terrain points, even if the terrain feature points are extracted as candidate points, the number of these candidate points is still very large. The use of serial k-means clustering method will cause the problem of low computational efficiency. Therefore, MPI parallel k-means algorithm is considered. When the clustering operation is completed, the viewpoints in each cluster should be sorted. When evaluating the quality of viewpoints, the sights of each viewpoint should be calculated. The calculation of the sights of a single viewpoint is time-consuming. Therefore, when calculating the viewshed of a large number of viewpoints, we adopt the OpenMP parallel program framework to reduce the calculation time. The last part is the filtering of candidate viewpoints. Considering that the viewpoints in the cluster needs to be re-sorted according to the quality of their viewing areas during the filtering process, OpenMP is also used for parallel calculation to improve the computational efficiency.

In summary, the parallelization method based on terrain point filtering algorithm is mainly divided into the following three parts: 1)parallelization of k-means clustering based on MPI; 2)parallelization of terrain point sorting based on OpenMP; and 3)parallelization of terrain point filtering based on OpenMP.

III. PARALLEL METHOD OF TERRAIN POINT FILTERING

A. Parallelization of k-means Clustering

MPI(Message Passing Interface) is one of the important technologies used in the cluster. It is an application program interface based on message passing model. The main feature is that between different processors, it uses the network to transfer messages to achieve mutual communication, and completes the synchronization between tasks. Message passing methods is generally used in distributed storage structure. Because the number of terrain points is very large, the original K-means clustering algorithm needs to spend a lot of time to calculate the Euclidean distance between each data point. After research, it can be found that the distance between each terrain point and the obtained terrain point as the clustering center is independent of each other and does not interfere with each other, so the algorithm can be calculated in parallel. Therefore, this paper first divides the original data into multiple sub-data blocks, and then allocates each sub-data block to an independent processor, and each processor only needs to calculate the distance between the topographic point in the assigned sub-data block and each cluster center separately, and then the clustering result of the sub-data block processed by each processor can be obtained. Finally, the global clustering results are obtained by calculating the output results of each processor through a method similar to the Reduce function in MapReduce framework. The proposal of the implementation of the parallelization of k-means clustering algorithm based on MPI is mainly to use master-slave mode. The task of the master node is to divide and distribute data. The slave nodes are allocated to with the local data to complete the calculation task and the results are fed back to the master node. The process of parallelization is as follows:

1)Assuming that the master node is process 0, the process reads data from the file and allocates the data to other processes.

2)The main process selects each cluster center and sends it to other processes.

3)Other processes calculate the distance between each point in the assigned data and the cluster center point, mark the category of each point, and then calculate the sum of the distances between each terrain point in each class and the cluster center, and finally return the results to the main process.

4)Process 0 calculates the new center point and sends it to other processes, and calculates the sum of distances between all terrain points in the class from other processes and its center point.

5)Repeat step 3 and 4 until the sum of the distances of all clusters in step 4 or the cluster center no longer change (i.e. convergence).

B. Parallelization of terrain viewpoint sorting

OpenMP is a thread-based programming framework, which is generally used in the parallel system of shared

memory. For a parallel program based on multithreading, it is necessary to block and schedule the loops in the program to better achieve load balancing, and ensure that the CPU utilization is the highest and has been in a busy state during most of the running time as long as possible. At the same time, it minimizes various costs, such as scheduling overhead, synchronization overhead and the switching overhead between contexts, so as to optimize performance. The parallelization of terrain point sorting in each cluster adopts the static balanced scheduling strategy. This part is to sort the quality of terrain points in each cluster. Each cluster does not interfere with each other, so the pragma instruction can be used to parallelize the sorting process. In the OpenMP-based terrain point sorting method, only clusters are partitioned. Each cluster is processed by a single thread, and there is no interference between clusters, so there is no shared variable. Therefore, the pseudo code of the parallel algorithm for sorting terrain point based on OpenMP is shown as follows.

```
input K clusters
#pragma omp parallel for num_threads(n)
(n is the specified number of threads)
for i from 0 to (k-1)
{
compute viewshed quality of every viewpoint;
sort these viewpoints according to their viewshed quality;
}
```

C. *Parallelization of terrain point filtering*

In the process of terrain point filtering, the terrain points in each filtering cluster may be moved to adjacent clusters, and each cluster has strong correlation and poor parallelism. When filtering terrain points, it is necessary to calculate the viewshed quality of the viewpoint in the adjacent clusters, which is also the most time-consuming part in the whole filtering process. Therefore, this part can use OpenMP to improve efficiency.

The calculation of viewshed quality involves the view and view repetition of a single viewpoint. When calculating the number of view repetitions, it is necessary to calculate the joint view of multiple viewpoints. The essence of this series of calculations is to calculate the viewshed of a single viewpoint, and the viewshed of a single viewpoint does not affect each other. Therefore, after calculating the single point viewshed with parallel computing mode, the main thread can calculate the viewshed quality.

First, the parallel for statement is placed before calculating the for loop in each view field. Parallel is used in combination with the for instruction, representing multiple threads to execute the code in the for loop in parallel. It also has two functions: generating parallel domains and assigning tasks. Since the viewpoint quality needs to calculate the joint view, and every time a field of view is calculated, the joint field of viewshed array V must be changed at the same time. Therefore, in the case of multi-threaded parallelism, the array V is a shared variable. Where the shared clause in OpenMP is used to modify V, the role of shared clause is to specify one or more variables as shared variables between multiple threads. Finally, after each thread completes the execution task, the viewpoint quality formula is usd to calculate its quality, because the time for each thread to complete the task will be different, and the barrier is generally used to synchronize the threads executing the code in the parallel region. When each thread reaches the barrier instruction, it needs to stop and wait until all other threads execute to the barrier to continue to execute the remaining part. Therefore, adding the #pragma omp barrier statement can realize thread synchronization. The pseudo code of the parallel calculation of viewpoint quality in the filtering process is shown as follows.

```
Input: cluster C; single viewpoint p; Initialize the joint view array V
# pragma omp parallel for shared（V）
for i from 1 to N (N is the number of viewpoints in cluster C)
{
 calculate the viewpoint of i;
 modify array V from the perspective of view i；
 #pragma omp barrier
}
compute viewshed Vp of viewpoint P;
compute the overage ratio O according to Vp and shared viewshed array V;
compute viewshed quality of viewpoint P according to O and Vp of P.
```

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experiment compares the serial method with the parallelization method, and verifies the computational efficiency of the three parallel methods.

1)Experiment with k-means based on MPI

The experimental environment is two computers with CPU@2.0GHz, and 4GB memory, and windows 10 system. A total of eight processes for the program are started to run. The data adopts 1000*1000 digital elevation model terrain with a resolution of 5 meters. In order to compare, 10000, 100000, and 1000000 data points were tested. The running time is shown in Table I.

Experimental result shows that the serial execution of K-means algorithm takes the longest time. After paralleling it with the MPI framework, the speed of data processing is significantly improved. With the continuous increase of the amount of data, the running time of serial algorithm rises sharply, but it also reflects the advantage of parallel computing. When the data sizes up to 10000, 100000 and 1000000, the running time of k-means based on MPI is about 1 / 2,1 / 10 and 1 / 433 of that of the serial k-means algorithm, and the speedups ratio reaches 2.0, 10.45 and 433.49 respectively. Thus, the efficiency of parallel computing algorithm proves to be greatly improved.

TABLE I. Comparison of running time and accelerate rate

Data size	10000	100000	1000000
k-means	149ms	7982ms	3652168ms
k-means based on MPI	75ms	764ms	8425ms
Accelerate rate	2.0	10.45	433.49

2)Experiment of terrain point sorting algorithm based on OpenMP

Since the OpenMP parallel program design framework is mainly to make full use of the CPU to accelerate the calculation, the experiment is carried out on a computer

with CPU@2.0GHz and 4GB memory, and windows 10 system. 1000 selected terrain feature points are clustered into 35, 45 and 65 classes as the experimental data. The number of threads can be specified with num_threads in OpenMP. The number of threads set in this experiment is 2 and 4, respectively. Table II shows the resulting experimental running time of 35, 45 and 65 classes sorted by serial and parallel methods.

It can be seen from table 2 that the use OpenMP to parallelize the sorting of terrain points in the cluster based on visual quality Index can improve the computational efficiency. Sorting viewpoints in each cluster requires many times of computations of the visibility of each viewpoint in the cluster, so serial sorting takes a long time. When sorting viewpoints in35, 45 and 65 clusters respectively, the computation time of serial algorithm is about 1.48, 1.52 and 1.55 times that of parallel computing with two threads, and 2.69, 2.91 and 2.8 times that of parallel computing with four threads. It can be seen that the utilization of the CPU is significantly increased when multithread parallel computing is enabled.

TABLE II. Comparison of running time

number of clusters	35	45	65
Serial algorithm	682s	649s	608s
Parallel algorithm with two threads	458s	425s	391s
Parallel algorithm with four threads	253s	223s	217s

3)Experiment on Topographic Point Filtering Algorithm Based on OpenMP

The experiment is also carried out on a computer with CPU@2.0GHz and 4GB memory, and windows 10 system. The experimental data are 35, 45 and 65 classes that originate from terrain feature points on D1 terrain after clustering and sorting operations applied. The number of threads is also 2 and 4. Table III shows the running time of serial candidate viewpoint filtering(CVF) and parallel CVF algorithms. It can be seen from Table III that using OpenMP to filter terrain points in parallel can significantly shorten the calculation time. When filtering terrain points, it takes the longest time to calculate the quality of viewpoints. In this experiment, 35, 45 and 65 clusters are filtered respectively. From the running time results, it is concluded that the time of serial computing is about 1.55, 1.27, 1.69 times of the parallel computing time when two threads are used, and 2.81, 2.8, 2.75 times of the computing time when using four threads.

TABLE III. Running time of serial CVF and parallel CVF

cluster number	35	45	65
serial CVF	45s	14s	22s
parallel CVF with two threads	29s	11s	13s
Parallel CVF with four threads	16s	5s	8s

V. CONCLUSION

In this paper, the rapid filtering method of candidate feature points with multiple viewpoints is parallelized step by step. The k-means algorithm is mainly parallelized by MPI framework, and using OpenMP is used to parallelize the sorting of terrain points in each cluster and the final terrain point filtering. Experiments show that the partial parallelization of computing time can significantly shorten the computing time and improve the computational efficiency. The parallelization method is carried out step by step, which is only suitable for candidate terrain point filtering method. The future research direction is to find a parallelization solution to make it suitable for various problems of multi-viewpoint visibility.

FUNDING

This work is supported by the National Natural Science Foundation of China [grant number 41771411].

REFERENCES

[1] Fisher P F. First Experiments in Viewshed Uncertainty: Simulating Fuzzy Viewsheds[J]. Photogrammetric Engineering and Remote Sensing, 1992, 58(3): 345-352.

[2] Pin L. Terrain Visibility Analysis Based on Line of Sight[J]. Computer Engineering and Applications, 2006, 42(8):223-226

[3] LEE J. Analyses of visibility sites on topographic surfaces[J]. International journal of geographical information systems, 1991, 5(4):413-429.

[4] Franklin, W.R., et al. Slope Accuracy and Path Planning on Compressed Terrain in Headway in Spatial Data Handling[C]. 13th International Symposium on Spatial Data Handling, Montpellier, France, 23-25 July, 2008.

[5] Bao S., Xiao N., Lai Z., Zhang H., & Kim C. Optimizing watchtower locations for forest fire monitoring using location models[J]. Fire Safety Journal, 2015, 71: 100-109.

[6] Akella M. R., Delmelle E. M., Batta R., Rogerson P. A., and Blatt A.. Adaptive Cell Tower Location Using Geostatistics[J]. Geographical Analysis, 2010, 42(3):227-244.

[7] Kammer F., Löffler M., Mutser P., et al. Practical Approaches to Partially Guarding a Polyhedral Terrain[J]. Geographic Information Science, 2014, 8728.

[8] Puppo E., & Marzano P. Discrete visibility problems and graph algorithms[J]. International Journal of Geographical Information Systems, 1997, 11(2), 139-161.

[9] Franklin W. R., Ray C. K., and Shashank M. Geometric Algorithms for Siting of Air Defense Missile Batteries[OL]. Research Project for Battle 2756, 1994. http://www. ecse. rpi. edu/Homepages/wrf/research/p/tec_report.pdf.

[10] Sorensen P. A. & Lanter D. Two algorithms for determining partial visibility and reducing data structure induced error in viewshed analysis[J]. Photogrammetric Engineering and Remote Sensing.1993,59. 1149-1160.

[11] Wang J., Robinson G. J., White K. A Fast Solution to Local Viewshed Computation Using Grid-Based Digital Elevation Models[J]. Photogrammetric Engineering and Remote Sensing, 1996, 62(10):1157-1164.

[12] Yu T., Xiong L., Gao M., Wang Z., et al. A new algorithm based on Region Partitioning for Filtering candidate viewpoints of a multiple viewshed[J]. International Journal of Geographical Information Science, 2016, 30(11), 2171-2187.

[13] Wang Y. W. & Dou W. F. A fast candidate viewpoints filtering algorithm for multiple viewshed site planning[J], International Journal of Geographical Information Science, 2020, 34(3): 448-463.

[14] Wu Y. L. An algorithm for computing viewsheds based on reference planes[J]. Wtusm Bulletin of Science and Technology, 2001, 1 (6):19-21.

[15] Kanungo T., Mount D. M., Netanyahu N. S., Piatko C. D., Silverman R., & Wu A. Y. An efficient k-means clustering algorithm: analysis and implementation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 24(7):881-892.

[16] Kim Y. H., Rana S., & Wise S. Exploring multiple viewshed analysis using terrain features and optimization techniques[J]. Computers & Geosciences, 2004, 30(9-10), 1019-1032.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

A Weight-based Clustering Routing Algorithm for Ad Hoc Networks

Zheng Sihai, Zheng Changrui
School of Artificial Intelligence
Jianghan University
Wuhan, China
E-mail: zhengsihai@jhun.edu.cn

***Abstract*—Clustering routing algorithm is an important part of network layer in Ad Hoc networks. Its main goal is to find an optimal multicast tree which can satisfy QoS constraints, the network can not only meet the QoS requirements of data services, but also improve the efficiency of limited network resources. WCRA(Weight-based Clustering Routing Algorithm), a clustering routing algorithm, is designed in this paper. It adopts the combination weighting method, and the cluster topology is flexible. If the clustering strategy is changed properly, the correlated cluster or uncorrelated cluster can be constructed quickly. It can also adjust itself according to the actual application, improve the fairness of node competition and optimize the cluster structure. Experiments show that it can significantly improve the transmission rate of mobile network nodes, enhance the stability of the network, and provide an application basis for QoS services in Ad Hoc networks.**

Keywords: Clustering algorithm; Ad Hoc Network; Node; Routing

Ⅰ. INTRODUCTION

Wireless mobile Ad Hoc network is a distributed network system without central control equipment. Every mobile node is both host and router. So far, the clustering routing algorithm has been researched a lot, and researcher had made many achievements.

CEDAR[1] is a clustered routing algorithm based on QoS. Its goal is to establish a virtual core structure for reliable routing information dissemination. The core area of CEDAR(Core Extraction Distributed Ad Hoc Routing Algorithm) is composed of core nodes. Any normal node must choose a core node as its manager. Each node performs periodic information exchange to determine and update the core area, trying to ensure the minimum number of nodes in the core area, and the core node selects an available path to cache in the routing table by broadcasting locally to neighboring nodes. The propagation distance of the state information depends on the bandwidth of the stability of the link. A stable link and sufficient bandwidth will enable more information to be found in the route search message. The advantages of CEDAR are mainly reflected in the discovery and maintenance of routes in the core area. When the scale of the network increases, the control overhead does not increase much, which is conducive to supporting the transmission of multimedia service streams.

ZHLS[2] is another layer-based routing protocol. ZHLS(Zone Based Hierarchical Link State Routing) defines two topological structures: node-level and interval-level. It contains two types of routing control messages: out-of-area messages and intra-area messages. Messages outside the zone are broadcast on the entire network to connect and communicate between the zones; messages within the zone are only broadcast within the zone, connecting nodes within the zone. Since it is only broadcast within the zone, it will not be affected by the movement of nodes in the zone. District-level status information has an impact.

All these classic algorithms have different degrees of limitations: some are only suitable for small-scale Ad Hoc networks with flat structure type; while some are only suitable for large-scale networks with hierarchical topology type; some never get the optimal problem solution in the formation of multicast trees, etc[3-4]. In this paper, a weight-based efficient clustering routing algorithm named WCRA (Weight-based Clustering Routing Algorithm) for Ad Hoc network is proposed. Simulation results show that its efficiency has been improved greatly.

In the rest of this paper, we first research on mathematical model in section II. Then, WCRA algorithm is analyzed in section III. The experiments and analysis are shown in section IV. At last, we offer conclusions in Section VI.

Ⅱ. WCRA ALGORITHM

WCRA is an efficient clustering algorithm based on weights for Ad Hoc network proposed in this paper.

A. Mathematical model

Suppose that some properties of Ad Hoc networks, such as processing power, wireless link stability, available bandwidth, and battery residual energy, can be represented by functions $f_1(x), f_2(x), \cdots, f_n(x)$. Then the composite function of the attributes of node $n_i \in V$ can be expressed as equation (1)

$$H(i) = \lambda_1 f_1(x) + \lambda_2 f_2(x) + \cdots + \lambda_n f_n(x) \quad (1)$$

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00038

Where λ_i is the weighting factor of the corresponding attribute. By taking different values of λ_i, we can simulate different application environment of Ad Hoc network. Using the normalized calculation method, the conditional constraint degree R_i of node $n_i \in V$ can be obtained, as shown in formula (2).

$$R_i = |H(i)|, 1 \le i \le n \quad (2)$$

The physical meaning of R_i is the communication capability of node n_i, and $0 \le R_i \le 1$. At the same time, it can be divided into four levels, representing four different communication states and capabilities, as shown in formula (3).

$$T_i = \begin{cases} 0, & 0 \le R_i < 0.25 \\ 1, & 2.5 \le R_i < 0.5 \\ 2, & 0.5 \le R_i < 0.75 \\ 3, & 0.75 \le R_i \le 1 \end{cases} \quad (3)$$

T_i=0 indicates that the communication capability of node n_i is weak; T_i=1 indicates that the communication capability is general; T_i=2 indicates that the communication capability is strong, and n_i can be used as a candidate node for the cluster head; T_i=3 indicates that the communication capability of the node is very strong and can be selected as cluster head.

B. Cluster head selection

WCRA algorithm uses four conditions to calculate the weight of nodes when selecting cluster head[5-7]. They are: the sum of the distance between the node and its neighbors, the deviation of the node degree and the residual energy of the node. As shown in Formula (4) .

$$W_i = \lambda_1 E_i + \lambda_2 M_i + \lambda_3 D_i + \lambda_4 \Delta_i \quad (4)$$

Where, W_i represents the weight of the node; λ_1、λ_2、λ_3、λ_4 respectively represent the coefficient of each weight, its size varies with different application environments, and is a system parameter; E_i represents the initial energy of the node; D_i is the sum of the distances between node n_i and each neighbor node; Δ_i is the absolute value of the difference between node n_i and the best node degree; M_i is the relative movement characteristic value of the node. Assuming that the mobility of node n_j relative to node n_i is $M_j(i)$, then its value can be calculated by formula (5).

$$M_j(i) = 10 \log \frac{New\_Power}{Old\_Power} \quad (5)$$

Where, *New_Power* is the strength of the wireless signal received by node n_j from node n_i, *Old_Power* is the intensity of the signal from n_i to n_j at the previous moment. If $M_j(i)$<0, it means that node n_j has left the communication range of node n_i, otherwise they are getting closer. The relative mobility of node n_i can be expressed by equation (6).

$$M_i = \frac{\sum_{j=1}^{k} |M_i(j)|}{k} \quad (6)$$

Where, k indicates that node n_i has k neighbor nodes. The larger the M_i is, the stronger the mobility of node n_i is, and vice versa.

The WCRA algorithm calculates the weight of candidate cluster head node according to equation (4). The node with the lowest weight will be selected as cluster head.

C. cluster size

If a cluster head manages more member nodes, it will consume its remaining energy too quickly and lead to the death of the cluster head, while the election of new cluster heads increases the control overhead of the network and initiates the re-run of the routing algorithm; in addition, too many cluster members will also cause congestion in the network. However, if there are too few member nodes in a cluster, it will lead to a waste of network resources. The following is an analysis of how the WCRA algorithm determines the number of member nodes in a cluster.

Assuming that in Ad Hoc network $G = \{V, E\}$, $n_i \in V$, where $i = 1, 2, \cdots, N$, BW is the total bandwidth of the wireless channel, then the maximum traffic *Flow* of each network node can be expressed as equation (7).

$$Flow = \frac{BW}{\sqrt{N}} \quad (7)$$

The cluster heads will form an upper layer backbone network among themselves, and each cluster head will have to communicate with both intra-cluster nodes and other cluster heads, so it must initiate multi-frequency communication with both inter-cluster and intra-cluster traffic.

Assuming that the number of member nodes in each cluster is the same, and there are m members in each cluster, there are N/m cluster heads. If the intra-cluster bandwidth is BW_1 and the inter-cluster bandwidth is BW_2, the traffic of each cluster member node can be expressed by equation (8), and the traffic on the cluster head can be expressed by equation (9).

$$Flow_{member} = \frac{BW_1}{\sqrt{\frac{N}{m}}} \quad (8)$$

$$Flow_{cluster} = \frac{BW_2}{\sqrt{m}} \quad (9)$$

The cluster head itself is also a member node in the cluster, so equation (10) holds.

$$\frac{m-1}{m} Flow_{member} \le Flow_{cluster} \quad 10)$$

The following analysis finds a value of m that maximizes $(m\text{-}1)/m \times Flow_{member}$, which is the optimal value for the number of cluster members determined by the WCRA algorithm.

Assuming that $m = m'$ is the value we need to solve, the upper limit of $(m\text{-}1)/m \times Flow_{member}$ can be expressed by equation (11).

$$F_{up\_member} = \frac{m'-1}{m'} \sqrt{\frac{8}{\pi}} \frac{BW_1}{\Delta} \sqrt{\frac{N}{m'}} \quad (11)$$

The upper limit of $Flow_{cluster}$ can be expressed by formula (12).

$$F_{up\_cluster} = \sqrt{\frac{8}{\pi}} \frac{BW_2}{\Delta} \sqrt{m'} \quad (12)$$

Obviously, if $F_{up\_member}=F_{up\_cluster}$, then equation (13) holds.

$$\frac{m'-1}{m'} \sqrt{\frac{8}{\pi}} \frac{BW_1}{\Delta} \sqrt{\frac{N}{m'}} = \sqrt{\frac{8}{\pi}} \frac{BW_2}{\Delta} \sqrt{m'} \quad (13)$$

From formula (13), formula (14) can be derived

$$m' = \frac{BW_1}{BW_2} \sqrt{N} + 1 \quad (14)$$

According to equation (14), the performance of the network can be optimized when the number of member nodes in the cluster is m'.

III. SIMULATION AND ANALYSIS

To evaluate WCRA and compare it to existing algorithm, simulations are performed. A Mobility Framework for NS2[8-9], a discrete event simulator written in C++, is used as a tool. the scene is 2000m×2500m, the number of nodes are respectively 50, 100, 150, 200, 250, 300, 350 and 400 in scene, and they are randomly distributed;

Fig. 1 shows data delivery rate is rapidly increased while the number of nodes is less than 200, but it is slowly decreased while the number of nodes is more than 200. That is because some paths will become longer if there are more nodes in network, more and more packets are dropped in process of transmission, then the two data delivery rates are decreased. With the increase of number of nodes, the network topologies become more and more complex. The packet delivery rate of WCRA is obviously higher than the other two algorithms by about 15%.

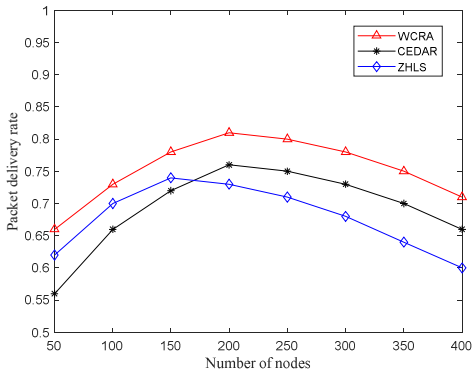


Fig. 1 Packet delivery rate

Fig. 2 shows the relationship between network lifetime and the number of nodes. With the increase of number of nodes, the network lifetime both incline slowly. That is because more nodes, more complicated topology and more data forwarded, which make load increase rapidly and residual energy is run out in a short time. Thus, network lifetime is declining. The network lifetime of WCRA is apparently longer than that of the other two algorithms in case of the same number of nodes.

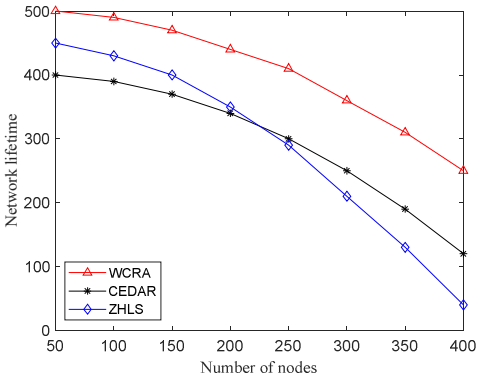


Fig. 2 Network lifetime

IV. CONCLUSION

This paper designs a clustering routing algorithm WCRA, which adopts a combined weighting method, and the determination of cluster topology is flexible. As long as the clustering strategy is appropriately changed, relevant or unrelated clusters can be constructed quickly. It can also make adaptive adjustments according to actual applications, improve the fairness of node competition, and optimize the clustering structure. Experiments show that it can significantly increase the transmission rate of mobile network nodes, enhance the stability of the network, and provide an application basis for QoS services in Ad Hoc networks.

ACKNOWLEDGMENT

This paper is supported by Provincial Teaching Research Project in Hubei Province of China(No: 2018308), Guiding Project of Science and Technology Research Plan of Hubei Provincial Department of Education of China(No: B2019241).

REFERENCES

[1] Xiaojie Liu; Ulrich Speidel. RAODV Routing Protocol for Congestion Detection and Relief in Ad Hoc Wireless Networks. International Journal of Interdisciplinary Telecommunications and Networking, v 13, n 4, p 21-34, October 1, 2021.

[2] Sarao, Pushpender. Ad Hoc On-Demand Multipath Distance Vector Based Routing in Ad-Hoc Networks. Wireless Personal Communications, v 114, n 4, p 2933-2953, October 1, 2020.

[3] Wu, Huazhu; Guo, Wenyou. Study on the routing algorithm of airport navigation lighting system based on mobile wireless ad hoc network. Journal of Computational Methods in Sciences and Engineering, v 20, n 4, p 1085-1096, 2020.

[4] Basomingera, Robert; Choi, Young-June. Learning from routing information for detecting routing misbehavior in ad hoc networks. Sensors (Switzerland), v 20, n 21, p 1-22, November 1, 2020.

[5] Vinayagam, J.K.; Balaswamy, C.H. Cross-layered-based adaptive secured routing and data transmission in MANET. International Journal of Mobile Network Design and Innovation, v 9, n 1, p 37-45, 2019.

[6] Sekar, P. Chandra; Mangalam, H. A power aware mechanism for energy efficient routing in MANET. International Journal of Networking and Virtual Organisations, v 21, n 1, p 3-18, 2019.

[7] Osman, M.; Syed-Yusof, S.. A survey of clustering algorithms for cognitive radio ad hoc networks. Wireless Networks, v 24, n 5, p 1451-1475, July 1, 2018.

[8] Marchang, Jims; Douglas, Roderick. Dynamic Neighbour Aware Power-controlled MAC for Multi-hop Ad-hoc networks. Ad Hoc Networks, v 75-76, p 119-134, June - July 2018.

[9] Shengli, Mao. Research on the application of NS2 network simulation based on clustering algorithm. Open Automation and Control Systems Journal, v 7, n 1, p 1210-1215, September 14, 2015.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

A Railway Station Scheduling Algorithm Based on Multi-Objective Particle Swarm Optimization Algorithm

RuiMin Deng,YongHua Li,JunYa Hu
School of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
Email: 546272518@qq.com

Qi Chen
Information Technology Department,
Chongqing Guoyuan Port Co, Ltd
Chongqing, China
Email: 358793710@qq.com

***Abstract*—In the railway station scheduling system under the water-rail combined transport mode, how to reasonably allocate equipment resources according to the operation tasks, improve the efficiency of mechanical loading and unloading, and reduce the cost of loading and unloading operations, has become an urgent problem in the process of building a modern port. In this paper, a railway station scheduling algorithm based on multi-objective particle swarm optimization algorithm (RSSAMOPSO) was designed. The multi-objective particle swarm algorithm was used to solve the multi-objective optimization problem of railway station scheduling. Aiming at the problems of decreasing convergence speed and the lost of the diversity of the solutions, a compression factor and an adaptive mutation operator are introduced. The improved multi-objective particle swarm algorithm (IM_MOPSO) is combined with the compression factor and adaptive mutation operator to obtain the optimal operation sequence and various equipment task allocation schemes. Experimental results show that the above method has a large practical significance and practical value for improving the efficiency of mechanical loading and unloading in the railway yard and reducing the operating cost of the railway yard.**

***Keywords*-railway station; particle swarm optimization; task scheduling; multi-objective;**

I. INTRODUCTION

As the hub of water and land transportation, the port undertakes 90% of the world's foreign trade products handling and transportation services.In order to realize the intellectualization of port railway station and yard, it is necessary to carry out the intelligent dispatching of railway station and yard under the mode of water-railway combined transportation[1].

The task scheduling problem has been proved to be a NP(Nonpolynomial) complete combinatorial optimization problem[1][2].Combinatorial optimization problems can be divided into single objective optimization and multi-objective optimization problems from the point of view of target objects.The task scheduling problem of railway station and yard is just a multi-objective optimization problem. The multi-objective optimization algorithm is actually seeking the balance point of mutual restriction of multiple optimization objectives.For complex multi-objective optimization problems, classical algorithms such as the main objective method, linear summation method, approximation method, etc.[3][4] have certain limitations, and they cannot be solved for those multi-objective problems that have conflicts or have no commonality.Particle swarm optimization algorithm[5] (Partical Swarm Optimization,PSO) convergence is fast, simple iteration, so the multi-objective particle swarm optimization (Multi - objective particle swarm optimization,Mopso) has become a meaningful direction to solve practical engineering problems[6].The difference between the MOPSO algorithm proposed by Coello and the particle swarm optimization algorithm is that the MOPSO algorithm introduces the concept that the external population can adapt to the grid mechanism, and the swarm particles and the value range of particles are varied[7].MOPSO has the following two innovations: 1) storing external populations through the use of external archives;2) Fast convergence and the introduction of a new mutation strategy.

II. RAILWAY STATION SCHEDULING ALGORITHM BASED ON MULTI-OBJECTIVE PARTICLE SWARM OPTIMIZATION ALGORITHM

In the process of railway station and yard dispatching, it is not only necessary to consider the operation sequence of fixed machinery, Inner container truck and frontal hoisting equipment, but also to ensure that there is no conflict among the equipment. The core of railway station and yard dispatching is the fixed machinery of railway station and yard. The operation sequence of fixed machinery of railway station and yard is determined first, and then the operation sequence of Inner container truck and frontal hoisting equipment is determined, as well as the destination of goods entering the port by train.

First, initialize the particle swarm.Then, the objective function is calculated, and the operation sequence of fixed machinery in railway station and yard is given according to the idea of table scheduling.The result of the objective function is taken as the filter object of the non-dominated solution, and the non-dominated solution is screened.Then calculate the following steps in a loop: 1) the crowding degree in external files and sort it in descending order;2) Select the global optimal;3) Update external files.

A. Objective function

The goal of railway station and yard dispatching under the mode of water-railway combined transportation is to minimize the time of task completion, the waiting time of equipment and the waiting time of ship.The total time to

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00039

complete a task refers to the time when the last device finishes the task, as shown in formula (1).The waiting time of equipment refers to the waiting time of different equipment during handover, including the waiting time of fixed machinery of railway station and yard and the waiting time of handover of Inner container truck. The waiting time of Inner container truck and frontal hoisting equipment handover is shown in Equation (2).The waiting time of the ship is shown in formula (3).

$$f_1 = \min(\max(hc_t, t \in T, t \neq t_{start})) \tag{1}$$

$$f_2 = \min(\sum_{t\in T}\sum_{c\in C} wc_t^c + \sum_{t\in T}\sum_{f\in F} wf_t^f + \sum_{t\in T}\sum_{R_s\in R} wR_t^{R_s}) \tag{2}$$

$$f_3 = \min(\sum_{t\in T_{sp}} gd_{T_{sp}} \times cv_l \times t_{tr}) \tag{3}$$

Among them, the specific calculation formula of the waiting time of each device is shown in (4) to (6).The calculation formula of the time for the fixed equipment of railway station and yard to arrive at the place where the task starts is shown in (7).The calculation formula for the time taken for the frontal hoisting equipment to travel to the handover area of the yard is shown in (8).

$$\begin{aligned} wc_t^c = \max(&\sum_{t,t_x\in T} R_{t,t_x}^s \times hx_t^{R_s} - \sum_{t,t_x\in T} c_{t,t_x} \\ &\times hx_t^c, 0) + \max(ht_t^f - \sum_{t,t_x\in T} c_{t,t_x} \times hx_t^c, 0) \end{aligned} \tag{4}$$

$$wf_t^f = \max(\sum_{t,t_x\in T} c_{t,t_x} \times ht_t^c - ht_t^f, 0) \tag{5}$$

$$R_t^{R_s} = \max(\sum_{t,t_x\in T} c_{t,t_x} \times hx_t^c - \sum_{t,t_x\in T} R_{t,t_x}^s \times hx_t^{R_s}, 0) \tag{6}$$

$$ht_t^f = \sum_{t,t_x\in T} f_{t,t_x} \times cp_{t_x}^f + \sum_{t,t_x\in T} f_{t,t_x} \times df_{t,t_x}^f \div fv_e \tag{7}$$

$$\begin{aligned} hx_t^{R_s} = &\sum_{t,t_x\in T} R_{t,t_x}^s \times cp_t^{R_s} + \sum_{t,t_x\in T} (1 - L_{t_x}) \\ &\times ds_t^{R_s} \times R_{t,t_x}^s \div rv_e \end{aligned} \tag{8}$$

B. Give the operating sequence of fixed machinery in railway yard

The loading and unloading task of railway station has a certain time sequence. DAG (directed acyclic graph) is more intuitive and easy to understand to describe the task scheduling. Therefore, DAG is used to describe the time sequence relationship of loading and unloading task of railway station.

The main objective of DAG-based method for assigning operation sequence of fixed machinery in railway station and yard is to properly assign tasks to fixed machinery in railway station and yard so as to minimize the execution time of all loading and unloading tasks. Task fitness TF_{mn} is defined to represent the similarity degree between the cargo types in task T_n and the set of all the remaining cargo types in the fixed machinery F_m of railway station and yard.

It is assumed that the number of goods types in the railway yard dispatching system is GT , the whole set of goods types is $\{1, 2, 3, ..., GT\}$, and the goods types in a task are expressed as $TG = (q_1, q_2, q_3, ..., q_{GT})$, where:

$$q_i = \begin{cases} 0, Goods\ i\ not\ in\ the\ task \\ 1, Goods\ i\ in\ the\ task \end{cases}, i \in [1, GT] \tag{9}$$

Similarly, the types of goods to be processed in real-time by fixed machinery in railway stations and yards in real time are expressed as $FG = (q_1, q_2, q_3, ..., q_{GT})$. At this time, the adaptability of task T_n corresponding to fixed machinery F_m in railway station and yard can be expressed as:

$$TF_{mn} = TG_n \times MG_m^T \tag{10}$$

Meanwhile, in order to prevent the task from starving to death,T_{ti} is set as the maximum acceptable processing time of the task, and the task processing priority of task T_n corresponding to the fixed machinery M_m of the railway station and yard is expressed as:

$$TP_{mnt} = \varepsilon \times \frac{TF_{mn}}{\left\| {TG_n}^T \right\|_1} + (1 - \varepsilon) \times \frac{dt - T_{Sys}}{T_{ti}} \tag{11}$$

Where,T_{Sys} represents the arrival time of the train at the port,ε is the weight coefficient, which can be adjusted according to on-site requirements, and dt represents the current time.

DAG task scheduling is the essence of the method of assigning the operation sequence of fixed machinery in railway station and yard.The classical algorithm of DAG task scheduling table scheduling algorithm.According to the characteristics of the problem, the table scheduling algorithm was combined with the improved multi-objective optimization algorithm, that is, the task height and the position of particles were taken as the priority of the task in the solution process.In the particle swarm, each particle has n dimensions, and each dimension corresponds to a task.The specific steps are as follows:

1) Initialization, each particle dimension corresponds to a loading and unloading task;

2) Sort in ascending order according to task height;

3) Within the same height range, sort by position in ascending order;

4) Determine whether the task list is empty, if it is empty, it will end, and if it is not empty, it will execute the next step;

5) Calculate the earliest start time $est(t_\xi, m_j)$ for assigning tasks T_i to the fixed machinery of each railway station according to formula (13);

6) Assign the task to the fixed machinery of the railway yard that minimizes the earliest start time of the task;

7) Delete the tasks that have been assigned to the machine, and repeat step 4;

8) Output the completion time of the last machine and the algorithm terminates.

C. Improved multi-objective particle swarm optimization algorithm

In order to improve the convergence speed and diversity, this paper makes the following improvements to

the standard multi-objective particle swarm optimization algorithm:

1)Using archives to save the found optimal solution,so as to realize elite preservation.

2)Compression factor is introduced to improve the speed update strategy to improve the convergence speed of the algorithm.

3)The idea of adaptive mutation was introduced to increase the diversity of particles.

4)The selection of global optimal solution is improved to further ensure the diversity of particles.

D. Particle velocity update strategy

The compression factor λ is introduced to improve the speed update, and the speed update strategy is shown in Formula (12).Compression factor λ as shown in formula (13).

$$V_{id}(t) = \lambda V_{id}(t-1) + c_1 rand_1(P_{id} - X_{id}(t-1)) + c_2 rand_2(P_{gd} - X_{id}(t-1)) \tag{12}$$

$$\lambda = \frac{2}{|2 - \varphi - \sqrt{\varphi^2 - 4\varphi}|} \tag{13}$$

$$\varphi = c_1 + c_2 \tag{14}$$

Where, t represents the current number of iterations; V_{id} represents the component of i particle in d dimensional velocity; g is the global optimal particle index number; P_{gd} is the component of all the particles in d dimension,which goes through the optimal position; X_{id} represents the position of i particle in d dimensions; P_{id} represents the best position of the i particle in the position experienced by itself in the d dimensional component; c_1 is individual learning factor, c_2 is social learning factor; $rand_1$ and $rand_2$ are random numbers between [0,1].

E. Updating and maintenance of external files

1) Design of external files: A set is designed to store the non-dominated solutions generated during previous iterations. This set is called the external archive.The idea of selecting the non-dominant solution is to select a particle from the particle swarm and judge the dominant relationship between this particle and other particles in the particle swarm. If no particle has a dominant relationship with the current particle is found, the current particle will be placed in the external archive set, that is, the current particle is the non-dominant solution.If a dominant particle is found in the particle swarm, the dominant particle is deleted from the particle swarm.Do this for the remaining particles in the current particle swarm until the set of particles in the swarm is empty.At this point, the solutions stored in external files are all the currently obtained non-dominated solutions[8].

2) Update external file set: In order to ensure the diversity of the optimal solution set, the selection and mutation operators in the genetic algorithm are selected to update and maintain the external file set.Using the roulette wheel selection method for several rounds of selection, the larger the crowded distance, the more likely it is to be selected, when search algorithm tends to be suspended, population may be trapped in local optimum, may also be close to the global optimal solution, in the particle's velocity approximate zero,the mutation operator can also help to update the position of the particle, especially in the case of population into a local optimum,the mutation operator plays a more important role,the mutation process is shown as follows.

$$\begin{aligned} child(x) &= p \times parent_1(x) + (1-p) \times parent_2(x) \\ child(v) &= \frac{parent_1(v)+parent_2(v)}{|parent_1(v)+parent_2(v)|} |parent_1(v)| \end{aligned} \tag{15}$$

$parent_1(v) parent_2(v)$ is the velocity of the parent particle; p is the probability of mutation.A possibly better external set can be obtained by replacing the dominant solution in the external file set by comparing the newly generated solution.In order to improve the global search ability of the algorithm, the adaptive mutation operation is adopted in the later period of operation, and the decreasing mutation probability is used in iteration to improve the convergence ability of the algorithm. The formula of the adaptive mutation probability is as follows.

$$p = 1 - \frac{t}{Iter} \tag{16}$$

Where, t is the current iteration number and $Iter$ is the maximum iteration number.

Therefore, the updating steps of external files are as follows: firstly, according to the congestion degree method, calculate the congestion degree of the solution in the file. Then, a probability value $b \in [0, 1]$ is generated randomly, and the value of b is compared with the mutation probability. If the value of b is small, the mutation operation is carried out. Mutation operation: sort the solutions according to the crowding degree, randomly select a non dominated solution with the crowding degree ranking in the first 50% as one of the mutation operators, randomly select a non dominated solution with the crowding degree ranking in the last 50% as another mutation operator, generate two solutions through cross mutation, and randomly select one of the two solutions to be stored in solution set D, Until all pairing and crossing are completed in the external file. Finally, prune the file and delete the redundant solution.

F. Select global optimal particle

The selection of the global optimal particle is the core of the multi-objective particle swarm optimization algorithm.When selecting the global optimal solution, the crowding distance of particles is arranged in descending order, and the optimal solution is selected randomly from the top 10% particles to further ensure the diversity and distribution of solutions.

G. The overall process of the IM_MOPSO

First initialize to form an initial population.Then, the objective function value of each particle is calculated according to each objective function, and the initial archive set A is obtained.Finally, the following five steps are executed in the loop until the maximum number of iterations is reached and the algorithm is finished.The five steps are

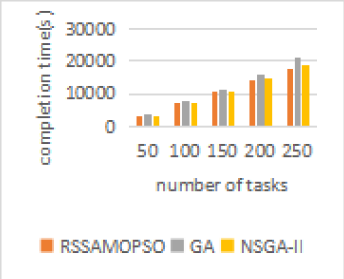


Figure 1. Completion time

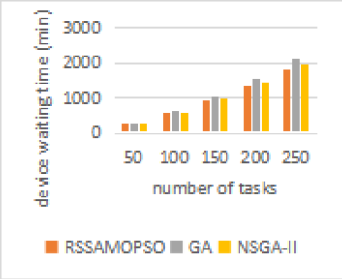


Figure 2. Waiting time of Equipment

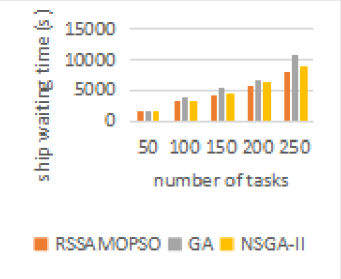


Figure 3. Waiting time of ship

as follows: 1) Calculate the crowding degree between each particle; 2) Sort the particles in descending order according to crowding degree; 3) Select one of the top 10% particles with crowding degree as the optimal solution; 4) Use formula (13) to update speed and position; 5) Calculate fitness values, update and maintain external files.

III. Experiments and Analysis

Traditional multi-objective genetic algorithm (NSGA_Π) is the classic algorithm to solve multi-objective optimization[9]. In order to verify the advantages and disadvantages of the RSSAMOPSO algorithm, the RSSAMOPSO algorithm is compared with the artificial experience, namely greedy algorithm (GA) ,and NSGA_Π.

50, 100, 150, 200 and 250 loading and unloading tasks are tested respectively, and the optimization results of each algorithm are compared. Table I shows the values of key parameters of RSSAMOPSO. The value is obtained by adjusting parameters through many experiments.

Results are shown in Figure 1 to Figure 3.When the number of tasks is greater than 150, the optimization target values of GA algorithm are obviously inferior to RSSAMOPSO algorithm and NSGA_Π algorithm.When the number of tasks is less than 200, there is little difference between the optimization objective results generated by RSSAMOPSO algorithm and NSGA_Π algorithm.When the number is greater than 200, the optimization target values generated by NSGA_Π algorithm scheduling are obviously inferior to RSSAMOPSO algorithm. It can be seen that RSSAMOPSO algorithm has better optimization ability compared with the other two algorithms when the number of railway station and field tasks is large.

IV. Conclusion

In order to improve the operation efficiency of port railway station and yard under the mode of water and railway combined transportation, this paper designs an intelligent scheduling method of railway station and yard, and uses multi-objective particle swarm optimization algorithm to solve the multi-objective optimization problem of railway station and yard dispatching.The results show that the algorithm in this paper has good optimization ability when the task quantity is large, which can effectively shorten the task completion time, ship waiting time and equipment waiting time, and improve the operation efficiency of railway stations and yards as well as the overall benefit of the port.

Table I
Key parameters of RSSAMOPSO

parameter	value
population size	200
external archive size	200
Inertia factor ω	0.4
learning factor c_1	2
learning factor c_2	2
Iterations	1000
weight of completion time ω_1	0.3
weight of device waiting time ω_2	0.2
weight of ship waiting time ω_3	0.5
Task processing priority weight ε	0.5

References

[1] Rajabi A, Saryazdi A K, Belfkih A, et al. Towards Smart Port: An Application of AIS Data[C]. 2018 IEEE 20th International Conference on High Performance Computing and Communications. Exeter, United Kingdom: IEEE, 2018:1414-1421.

[2] Yau K, Peng S, Qadir J, et al. Towards Smart Port Infrastructures: Enhancing Port Activities using Information and Communications Technology[J]. IEEE Access, 2020, 12(99):1-1.

[3] Li He,Qinghuai Liu. Multi objective optimization theory and continuous method[M]. Science Publishing Company, 2015.

[4] Baolin Chen. Optimization Theory and Algorithm[M]. Tsinghua University Press, 2005.

[5] Kennedy J, Barnhart R C. Particle Swarm Optimization[C].1995 IEEE International Conference on Neural Networks, 1995:1942-1948.

[6] Zhang X W, Liu H, Tu L P. A modified particle swarm optimization for multimodal multi-objective optimization[J]. Engineering Applications of Artificial Intelligence, 2020, 95:3-5.

[7] Coello C A C, Pulido G T, Lechuga M S. Handling multiple objectives with particle swarm optimization [J]. IEEE Transactions on Evolutionary Computation, 2004, 8(3):256-279.

[8] Yulong Xu, Xu Pan, Zhongyi Wang, et al. Efficient three objective differential evolution method based on Pareto non dominance[J]. Computer application research, 2019, 329(03):184-190+195.

[9] Jun Guo. Optimization of non dominated sorting genetic algorithm with elitist strategy[D]. Liaoning University,2017.

Recognition of power quality disturbance based on artificial bee colony algorithm to optimize kernel extreme learning machine

QU Li-ping*
College of Electrical and Information Engineering,
Beihua University
Jilin, China
E-mail: 309637683@qq.com

HE Chang-long，ZHANG Jie，GAO Tai-lu
College of Electrical and Information Engineering,
Beihua University
Jilin, China
E-mail: 13147776280@163.com

Abstract—**Aiming at the problem of classification and recognition of power quality disturbance signals, this paper proposes an artificial bee colony algorithm to optimize the classifier model of the kernel extreme learning machine. First, the wavelet transform is used to extract the features of the power quality disturbance signal simulated by Matlab, and then the traditional back-propagation (BP) neural network feedforward neural network is used for classification and recognition. Due to the inherent limitations of the algorithm, the network training calculation is large, the calculation is complicated, and the time is long. Disadvantages, and the proposal of the extreme learning machine just solves this problem, but the accuracy of the classification cannot meet the requirements, and the network needs to be optimized. Therefore, this paper proposes to use the artificial bee colony algorithm to optimize the nuclear extreme learning machine, and classify and recognize the disturbance signal under the optimal parameters of the nuclear extreme learning machine. The simulation results show that after the artificial bee colony algorithm is optimized, the correct rate of classification and recognition has increased by nearly 20%, and the misjudgment rate has dropped to about 3%.**

Keywords- Power quality; wavelet transform; nuclear extreme learning machine; artificial bee colony algorithm

I. INTRODUCTION

In recent years, with the continuous advancement of the industrialization process, the increasing power consumption and the switching of a large number of non-linear and impact loads have made power quality problems more and more serious, especially for users who use high-precision instruments. Putting forward higher requirements for power quality, and providing users with high-quality power is also the development trend of the power supply sector. In order to provide users with high-quality power, first of all, it is necessary to accurately identify various disturbance signals in the power, and take corresponding compensation measures for different disturbances; therefore, accurately identifying the characteristics of power quality disturbance signals is a prerequisite for improving power quality.

At present, the commonly used feature vector extraction methods for power quality disturbance signals include Fourier Transform (FT), Short-Time Fourier Transform (STFT), Wavelet Transform (WT), Hilbert Huang Transform (HHT), etc. It is introduced that FT is integrated over the entire time period without specific time information, resulting in a large error in the spectral distribution of abrupt signals. It has good analysis capabilities for stationary signals, but for unstable electrical energy. The quality disturbance signal cannot be detected and analyzed. In order to overcome this shortcoming, STFT is introduced on the basis of it, that is, time window is added to FT, and the non-stationary signal is regarded as a series of short-term stationary signals. Because the width of the window function is fixed, it is difficult to determine the width of the window function that matches the signal for unknown signals. In order to solve this problem, WT is adopted, which is based on the idea of STFT localization, which can flexibly change the time window and frequency window. Through multi-scale analysis, it can better extract the characteristics of non-stationary signals and better retain the signals. The difference between the two, it is convenient to improve the accuracy of classification. For the method of feature recognition, neural networks are currently used more frequently. BP is used to classify and recognize disturbance signals, and the network uses gradient descent method for iterative training[7]. In the process, it is necessary to constantly adjust the weights, which leads to long training time and easy to fall into local minimums. In order to solve this problem, in 2005, Professor Huang Guangbin of Nanyang Ligong University proposed the concept of an extreme learning machine in the paper [1]. For a single hidden layer feedforward neural network, the input layer weight w and threshold b are randomly performed assignment, the output layer weight β is obtained by the generalized inverse matrix theory, the training speed is greatly reduced, and the generalization performance is good. Professor Huang Guangbin has proved its feasibility in the literature [1].

Therefore, this paper uses wavelet transform to extract the characteristic vector of the disturbance signal, and uses the extreme learning machine and the nuclear extreme learning machine optimized by the artificial bee colony algorithm to classify and recognize, carry out simulation experiments.

II. FEATURE EXTRACTION OF POWER QUALITY DISTURBANCE SIGNAL

A. Typical power quality disturbance model

According to the relevant international power quality disturbance standards, the power quality disturbances in the actual power system are concentratedly reflected as transient voltage quality disturbances: voltage swells, voltage sags, voltage interruptions, transient oscillations, notches, flicker, and harmonics, these seven kinds of waves [4]. These disturbances are short in duration and random in nature. Other types of power quality disturbances are less likely to occur and have little harm.

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00040

Therefore, this article takes these seven power quality disturbances as examples for analysis and research.

Perform Matlab simulation on the above 7 kinds of disturbance signals and a standard signal, as shown in Figure 1 below, where *f*=50HZ, *T*=0.02s, 100 points are sampled in each cycle, 10 cycles, 1000 sampling points are sampled.

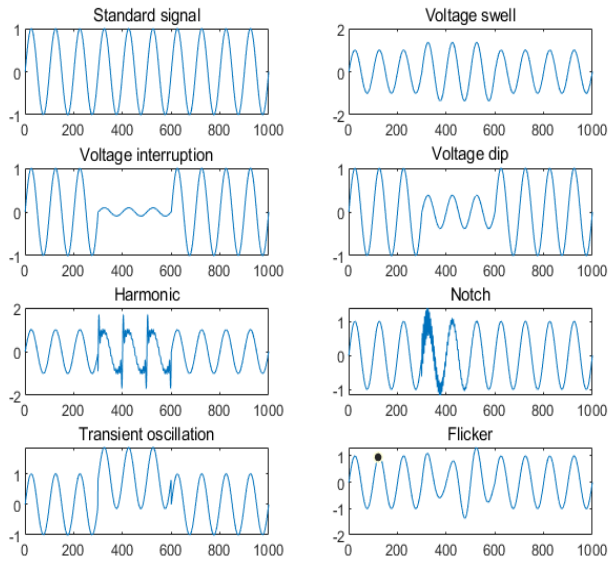


Fig. 1 Disturbance signal waveform

B. *Principle of wavelet transform*

Wavelet transform is a powerful mathematical transform that analyzes the time-frequency characteristics of a signal, and its principle is similar to that of Fourier. Wavelet transform is a wave function of a specific small area. It is translated, stretched, and superimposed to replace the original signal. Its essence is to measure the local similarity between the analyzed signal waveform and the wavelet waveform used. The higher the similarity, the larger the corresponding wavelet coefficient. The essence of wavelet transform is to express a signal function using wavelet function and wavelet transform coefficients. For the time domain signal $f(x)$, its wavelet transform can be expressed as:

$$f(x) = \sum s_{a,b} \Psi_{a,b} \quad (1)$$

Among them, $s_{a,b}$ is the wavelet coefficient and $\Psi_{a,b}$ is the wavelet function. For function $\Psi(t) \in L^2(R)$, if its Fourier transform satisfies:

$$C_\Psi = \int_{-\infty}^{+\infty} \frac{|\Psi(w)|^2}{|w|} dw < \infty \quad (2)$$

It is said that $\Psi(t)$is a basic wavelet. Common wavelets include Haar, Biorthogonal, Coiflet, Daubechies, etc. After the mother wavelet is translated and stretched, a wavelet sequence family $\{\Psi_{a,b}\}$ is obtained:

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}} \Psi\left(\frac{t-b}{a}\right), a > 0, b \in R \quad (3)$$

In the formula: a——scale factor b——translation factor;

For the continuous wavelet transform formula of any function $f(x)$:

$$W(a,b)_f = \langle f, \Psi_{a,b} \rangle = \frac{1}{\sqrt{|a|}} \int f(t)\, \Psi^*\left(\frac{t-b}{a}\right) dt \quad (4)$$

In the formula, when a and b take continuous values, it is continuous wavelet transform, which is mainly used in theoretical research. In practical applications, a and b are discretized, and the discrete wavelet transform is:

$$W(a,b)_f = a_0^{-m/2} \int_{-\infty}^{+\infty} f(t) \Psi_{m,n}^*(a_0^{-m} t - n b_0) dt \quad (5)$$

In the formula, $a = a_0^m$ ； $b = n b_0 a_0^m$ ； m, n are integers; a_0 is a scaling step greater than 1; $b_0 > 0$ and is related to the mother wavelet form.

C. *Equations*

By performing wavelet transform on the power quality disturbance signal, the wavelet transform coefficients obtained have the effective characteristics of the disturbance signal. Because the wavelet coefficients are long and complex and cannot intuitively characterize the characteristics of the signal, it is possible to perform certain operations on these coefficients. On the premise of keeping the original signal features, try to reduce the dimensionality of the feature vector, so as to reduce the size of the input vector when these features are applied to the classification neural network.

Using MATLAB to generate 200 samples of 7 common power quality disturbance signals and a standard signal, 1600 samples in total are abtained simulate the actual situation and ensure the reliability of the analysis results, the parameters of each disturbance (such as the start and end time, amplitude, duration, etc.) of each disturbance change randomly within an allowable range. The sample data is obtained, and the appropriate wavelet basis function is selected for wavelet transformation. Comprehensively considering the characteristics of the disturbance signal, this paper uses the db4 wavelet to decompose the disturbance signal in 8 layers, and the sum of the squares of the wavelet coefficients of each layer will be obtained as its energy representation [5]:

$$E_i = \sum |x_j(k)|^2 \quad (j = 1,2,\ldots 10) \quad (6)$$

Select the standardized energy feature P as the new feature input, and the standardized formula is as follows:

$$P = [p_1, p_2, \ldots p_{10}] \quad (7)$$

$$p_i = \sqrt{\frac{E_j}{k}} \quad (8)$$

In the formula, $j = 1,2,\ldots,10$; k is the number of sampling points. Then the data is normalized between [0,1] to improve the accuracy of distinguishing between data. Figure 2 shows the energy characteristics corresponding to 10 scales of normal signals and 7 kinds of disturbance signals. It can be seen intuitively that the energy of wavelet coefficients is mainly concentrated on scales 4, 5, and 6, when it reaches scale 1and10, it almost drops to 0, and then The next decomposition loses the distinguishing degree of features. If the energies of scales 1 and 10 are also used as the eigenvectors of the disturbance signal, it will not only increase the amount of calculation, but at the same time have little effect on the discrimination between the data, and the gain is not worth the loss. Therefore, this paper chooses scale 2 to scale 9 as the 8-dimensional feature vector of the disturbance signal, which not only preserves the signal characteristics, but also reduces the computational complexity.

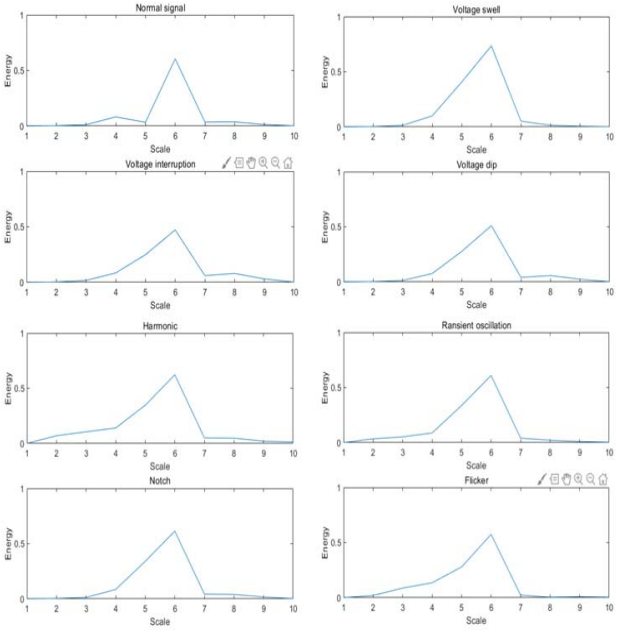


Fig. 2 Disturbance signal energy distribution diagram

III. CLASSIFICATION AND RECOGNITION OF DISTURBANCE SIGNALS

A. BP neural network

The BP (Back Propagation) network was proposed in 1986 by a scientific group headed by Rumelahart and McCelland. It is a multi-layer feedforward neural network trained on the error back propagation algorithm. It is currently one of the most widely used neural network models.

The 1600 sets of sample signals extracted above are divided into training set and test set, with 1200 as training and 400 as testing. There are 400 test samples, 50 samples for each type of signal, and the classification accuracy of each type of signal obtained through simulation is shown in Table 1 below. It can be seen that the training time of the BP neural network is relatively long and the convergence speed is slow. In the actual power system, for the massive detection data, due to the inability to quickly feedback, it may cause a major accident in the power grid.

TABLE 1 BP CLASSIFICATION RESULT

Disturbance type	test samples	Correct numbers	Correct rate(%)	Total correct rate (%)	Training time (s)
Standard signal	50	50	100	76.3	23
Voltage swell	50	47	94		
Voltage dip	50	50	100		
Voltage interruption	50	11	22		
Harmonic	50	46	92		
Transient oscillation	50	6	12		
Notch	50	50	100		
Flicker	50	45	90		

B. Kernel Extreme Learning Machine

In 2005, Professor Huang Guangbin proposed the ELM concept for a single hidden layer feedforward neural network, which is to randomly generate the connection weights between the input layer and the hidden layer. The training process does not need to be adjusted, and the optimal solution can be obtained by only setting the number of neurons in the hidden layer [6].

A single hidden layer neural network with L neurons in the hidden layer can be expressed as:

$$\sum_{i=1}^{L}\beta_i \cdot g(w_i \times X_i + b_i) = O_j, j = 1,2,\dots,n; \quad (9)$$

When the number of samples is equal to the number of hidden nodes, the output can be approximated with zero error. Given N different samples (X_i, Y_i), where X is an arbitrarily differentiable activation function $g(x)$, for an Single hidden layer feedforward neural network, w_i and b_i are arbitrarily assigned, and the output matrix H of its hidden layer is invertible , And$\|H\beta - Y\| = 0$[7].

In this regard, Professor Huang Guangbin, based on the extreme learning machine, draws on the concept of the kernel function in the support vector machine, and proposed the kernel extreme learning machine algorithm, which greatly improves the generalization ability and stability of ELM.

Randomly select samples of length N, $(X_i, t_i) \in R^N \times R^M, i = 1,\ 2,\ \cdots,\ N,\ X_i \in R^N$as input. Its expression is:

$$f(X_i) = h(X_i)H^T(HH^T + I/C)^{-1}T \quad (10)$$

In the formula, $h(X_i) = G(w_i X_i + b_i)$ is the output row vector of the hidden layer, G is the activation function, HH^T is the inner product form of $h(X_i)$, which can be replaced by the kernel matrix Ω_{KELM}:

$$\Omega_{KELM} = h(X_i)h(X_i)^T = K(X_i X_j) \quad (11)$$

Therefore, the solution formula of KELM is as follows:

$$f(X) = \begin{bmatrix} K(X, X_1) \\ \vdots \\ K(X, X_N) \end{bmatrix}^T (I/C + \Omega_{KELM})^{-1}T \quad (12)$$

Among them, $\beta = (I/C + \Omega_{KELM})^{-1}T$ is the output weight. There are a variety of kernel functions to choose from in KELM, such as linear kernel function (linear), polynomial kernel function (polynomial), Gaussian radial basis kernel function (RBF), etc. This article selects the RBF kernel function and defines it as follows:

$$K(X_i,\ X_j) = exp(-\frac{\|x_i - x_j\|}{\gamma^2}) \quad (13)$$

Although KELM introduces a kernel function and a penalty factor C to solve the problem of random initialization of ELM's input weights and obtains a stable predictive output. But its performance is easily affected by the penalty coefficient C and the kernel parameter γ.

C. Artificial bee colony algorithm optimizes nuclear extreme learning machine

Artificial bee colony (ABC) was proposed by Turkish scholar Karaboga in 2005. It simulates the honey-collecting behavior of bees to solve some multi-dimensional and multi-mode optimization problems in life. It was originally applied to numerical optimization

problems. Since its proposal, it has received great attention from many scholars, and has been widely used in many fields such as neural networks, data mining, and image recognition.

In this paper, the ABC algorithm and the kernel extreme learning machine are combined, and the ABC algorithm is used to find the optimal parameters of the kernel function to the network to achieve better generalization mapping ability [8]. The kernel function in this article selects the Gaussian kernel function, because it not only has strong learning ability, but also has better generalization, which can achieve good results in practical applications [10].

1) Create a KELM network.
2) Initialize the parameters of the ABC algorithm. Including the size of the colony (N_c), the number of bees (N_e), the number of following bees (N_o), the number of solutions (N_s), the limit value (limit), the maximum number of cycles (MNC) and the initial D dimension Solution $X_i(i = 1,2\cdots N_s)$. Among them, N_c, N_e, N_o, N_s satisfy the following relationship:

$$N_c = 2N_s = N_e + N_o, N_e = N_o \quad (14)$$

The D-dimensional solution vector X_i represents the two parameters of the Gaussian kernel function, and the initial solution is a randomly generated value between (-1, 1).

3) Calculate the fitness of each solution according to the following formula.

$$f(X_i) = \begin{cases} 1 & MSE_i = 0 \\ \dfrac{1}{MSE_i + 1} & MSE_i > 0 \end{cases} \quad (15)$$

In the formula, MSE_i represents the network mean square error of the i-th solution.

4) The honeybee searches for a new solution based on the current memory solution.

$$V_{ij} = X_{ij} + rand(-1,1)(X_{ij} - X_{kj}) \quad (16)$$

5) Calculate the possible value P_i of each solution.

$$P_i = \frac{f(X_i)}{\sum_{n=1}^{N_s} f(X_i)} \quad (17)$$

In the formula, $f(X_i)$ is the fitness value of the i-th solution, and the follower bee searches for a new solution from the neighborhood of the existing solution according to these possible values.

6) If the number of failures to update the solution X_i exceeds the preset limit value, then this solution cannot be optimized, it needs to be discarded, and then a new solution is generated to replace it.

$$X_i = X_{min} + \text{rand}(0,1)(X_{max} - X_{min}) \quad (18)$$

7) If the number of iterations is greater than the maximum number of cycles MCN, the training ends, otherwise, return to step 4.
8) Input the optimized optimal solution into the network, and use the data to simulate and test the network.

D. Experimental results and analysis

The performance of ELM and ABC-KELM was tested and compared on the Matlab 2018b simulation platform, and the experimental results were used to verify whether the ABC optimized nuclear extreme learning can improve the accuracy of classification. The experiment has a total of 1600 sets of sample data, 1200 sets are used as training samples, and 400 sets are test samples. The test samples include: seven power quality disturbance signals and normal signals, each with 50 samples. The parameters of the ABC algorithm are set in the experiment: the colony size is 200, the maximum limit number is 50, and the maximum number of cycles (MNC) is 100. According to the modeling steps in 2.3, the number of bees are both 100, that is, $N_s = N_e = N_O = 100$, with the mean square error (MSE) as the performance standard, the search range of the penalty factor C and the kernel parameter γ are [10,1000] and [0.01,1], respectively. In order to avoid the impact of sample differences on the experiment, the same training sample and test sample are used as the training set and test set of ELM and ABC-KELM. The simulation results are as follows:

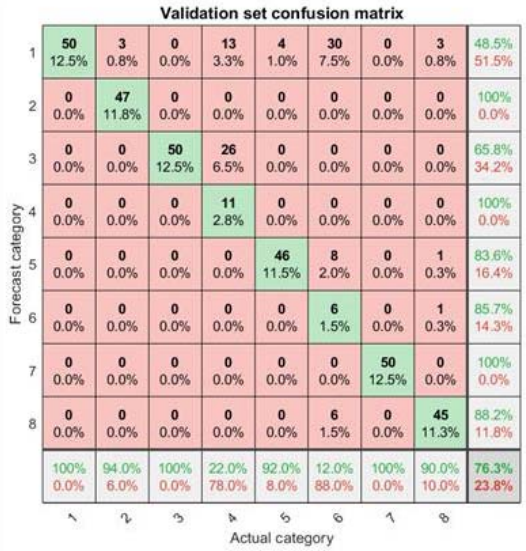


Fig. 4 Validation set of ELM

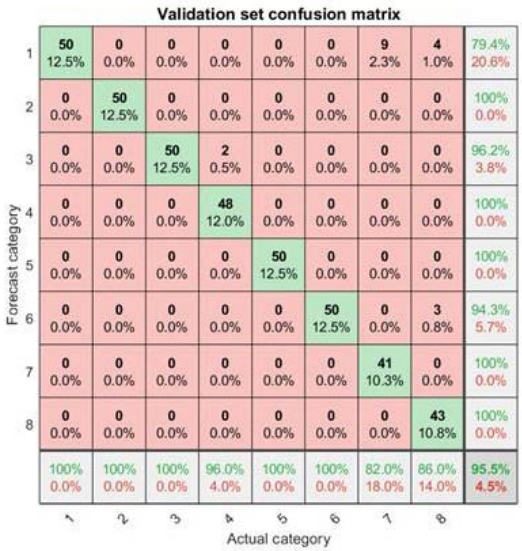


Fig. 5 Validation set of ABC-KELM

It can be clearly seen from the above results that inputting the same sample and using the extreme learning machine for classification and recognition, the poorer classification is the fourth type (voltage sag), 50 test samples, only 11 samples are accurately identified. The correct rate is only 22%. In addition, 13 samples were incorrectly classified into the first category (standard signal), 26 samples were incorrectly classified into the third category (voltage interruption), and in addition, The sixth type (transient oscillation) does not work well. Of the 50 samples, only 14 were accurately identified, the accuracy rate was less than 30%, 26 samples were incorrectly classified into the first class (standard signal), and 10 samples were incorrectly classified. To the fifth category (harmonic), this classification effect is far from reaching the requirements of engineering applications. In order to solve this problem, an artificial bee colony algorithm is proposed to optimize the kernel extreme learning machine to improve the classification accuracy. Through the iterative optimization of the bee colony, the optimal fitness solution of the network is found, and an optimal solution based on the mean square error is obtained. It can be seen intuitively from Figure 5 that the accuracy rate of the fourth category (voltage sag) is 96%, and only 2 samples are incorrectly classified into the third category (voltage interruption) and the sixth category. It has achieved the full recognition effect, and the total accuracy rate has reached 95.5%.

From the table, it can be seen that the artificial bee colony algorithm optimizes the kernel extreme learning machine very well. The accuracy of classification and recognition has increased by about 20%, and the classification accuracy has been greatly improved. However, in actual engineering applications, not only the classification accuracy must be considered, but also the classification speed must be considered to meet actual needs.

TABLE 4 COMPARISON OF ELM AND ABC-KELM CLASSIFICATION

Classification	Training samples	Test sample	Average accuracy
ELM	1200	400	7[illegible]. 7%
ABC-KELM	1200	400	97. 2%

IV. CONCLUSION

This paper proposes a power quality disturbance classification method based on the ABC algorithm to optimize the kernel extreme learning machine. First, db4 wavelet is used to decompose common power quality disturbance signals in 8 layers, and the square sum of wavelet coefficients is used to characterize its energy characteristics as the characteristic vector of the disturbance signal; then the traditional BP feedforward neural network is used for classification and recognition, due to its algorithm Inherent limitations, the network training calculation is large, the calculation is complicated, the time is long and other shortcomings, and the proposal of the extreme learning machine just solves this problem, but the accuracy of the classification does not meet the requirements, and the network needs to be optimized. Therefore, the ABC algorithm is used to optimize the nuclear extreme learning machine, and the kernel extreme learning machine under the optimal parameters is used to classify the disturbance signal. After simulation experiments, it is verified that this method has a great improvement in the classification effect of the traditional extreme learning machine. It has certain engineering application value.

REFERENCES:

[1] Das Soumya Ranjan,Ray Prakash K.,Sahoo Arun K.,Singh Krishna Kant,Dhiman Gaurav,Singh Akansha. Artificial intelligence based grid connected inverters for power quality improvement in smart grid applications[J]. Computers and Electrical Engineering,2021,93.

[2] Kumar Ashish,Tomar Harshit,Mehla Virender Kumar,Komaragiri Rama,Kumar Manjeet. Stationary wavelet transform based ECG signal denoising method.[J]. ISA transactions,2021,114.

[3] Gong Jie,Li Dayi,Wang Tingkang,Pan Wenhao,Ding Xinzhi. A comprehensive review of improving power quality using active power filters[J]. Electric Power Systems Research,2021,199.

[4] Guo Shifeng,Feng Haowen,Fenga Wei,Lv Gaolong,Chen Dan,Liu Yanjun,Wu Xinyu. Automatic Quantification Of Subsurface Defects By Analyzing Laser Ultrasonic Signals Using Convolutional Neural Networks And Wavelet Transform.[J]. IEEE transactions on ultrasonics, ferroelectrics, and frequency control,2021,PP.

[5] S. Syed Rafiammal,D. Najumnissa Jamal,S. Kaja Mohideen. Detection of Epilepsy Seizure in Adults Using Discrete Wavelet Transform and Cluster Nearest Neighborhood Classifier[J]. Iranian Journal of Science and Technology, Transactions of Electrical Engineering,2021(prepublish).

[6] Ji Xiu,Wang Hui,He Lin. New energy grid-connected power quality management system based on internet of things[J]. Sustainable Computing: Informatics and Systems,2021,30.

[7] Zhou Feifei,Yang Hao,Wang Jiahe,Sun Xi'an,Wu Xu. Prediction of the Quantity of Boxed Meals Preparation by High-Speed Train Based on ELM[J]. Journal of Physics: Conference Series,2021,1910(1).

[8] Mrutyunjaya Sahani,P.K. Dash,Debashisa Samal. A real-time power quality events recognition using variational mode decomposition and online-sequential extreme learning machine[J]. Measurement,2020,157.

[9] Junchang Xin,Zhongyang Wang,Shuo Tian,Zhiqiong Wang. NMR image segmentation based on Unsupervised Extreme Learning Machine[J]. Multidimensional Systems and Signal Processing,2017,28(3).

[10] Linqi Zhu,Chong Zhang,Yang Wei,Xueqing Zhou,Yuyang Huang,Chaomo Zhang. Inversion of the permeability of a tight gas reservoir with the combination of a deep Boltzmann kernel extreme learning machine and nuclear magnetic resonance logging transverse relaxation time spectrum data[J]. Society of Exploration Geophysicists and American Association of Petroleum Geologists,2017,5(3).

Improved Yolov4-tiny Algorithm Based on Cascade Residual Dilated Fusion

Qinggang Gong
School of Computer Science and Artificial Intelligence
Wuhan University of Technology
Wuhan 430063, China
e-mail: 517422877@qq.com

Yefu Wu*
School of Computer Science and Artificial Intelligence
Wuhan University of Technology
Wuhan 430063, China
e-mail: wuyefu1988@qq.com

***Abstract*—Performance of object feature extraction and fusion is not high in the detection process of yolov4-tiny algorithm, which leads to the low accuracy. An improved algorithm, dcca-yolov4-tiny based on cascaded residual dilation fusion mechanism embedded in location information, is proposed. The dilated convolution is used to increase the receptive field, and based on the idea of residual, the cascade residual fusion mechanism is used for feature fusion to improve the effect. The spatial features containing position information are embedded into the dilation convolution, so that the network can adaptively learn the spatial feature information. Two fusion methods are used combined with the position information to further improve the fusion effect of different object receptive field features and improve the accuracy of object detection. Experimental results on the VOC(Visual Object Classes) test set show that the proposed cascade residual dilated fusion module can promote the algorithm to effectively improve the accuracy of object detection. The mAP can reach 80.86%, which is 4.73% higher than the original model.**

Keywords—cascaded residual dilated convolution, position information, feature fusion, receptive field

I. INTRODUCTION

With the rapid improvement of computer hardware, object detection based on deep learning has made great progress in speed and accuracy, and is widely used in Unmanned Aerial Vehicle(UAV), vehicle detection and Intelligent Transportation System(ITS).

At present, a variety of improvement schemes are proposed for the accuracy and speed of target detection. Lin et al. [1] fused the shallow features with rich spatial information and the deep features with rich semantic information, on the premise of adding a small amount of calculation, effectively fused the shallow and deep multi-scale features to improve the accuracy of object detection. This idea of multi-scale feature fusion is reflected in documents [2,3]. Liu et al. [4] uses the human visual mechanism and uses parallel dilated convolution to increase the receptive field to form a branch structure, which effectively improves the object detection accuracy. Fu et al. [5] used deconvolution in SSD(Single Shot multibox Detector) to fuse shallow and deep features in a top-down manner, thereby improving the accuracy of object detection. Xu et al. [6] improved SSD algorithm by using lightweight network and applied it to helmet wearing detection, which improved the object detection speed with less loss of accuracy. Gao et al. [7] effectively improved the object detection accuracy by introducing the attention mechanism into CenterNet and combined with the auxiliary detection module. Redmon et al. [8] proposed yolo(you only look once) object detection algorithm to realize real-time detection.

II. CASCADE RESIDUAL DILATED FUSION MECHANISM

Through research on the above improved scheme, a cascaded residual dilated fusion module embedded with location information is proposed to improve the detection performance of yolov4-tiny which is designed based on yolov4 [9].

A. Dilated Convolution

Dilated convolution [10] uses different dilated factors, which increases the receptive field when the convolution kernel size is the same. In formula 1, K_{m+1} is the size of the current convolution kernel, RF_m and RF_{m+1} are the receptive fields of the previous layer and the current layer respectively, and S is the product of the stride of the previous m layer.

$$\mathrm{RF}_{m+1} = \mathrm{RF}_m + (K_{m+1} - 1) \times S \qquad (1)$$

Because the convolution kernel is expanded by filling the convolution gap with zero value, the discontinuity of convolution characteristic information may be caused. As shown in Fig 1, this paper uses dilated convolution with dilated factors of 1, 2, 3 for feature superposition to reduce feature discontinuity.

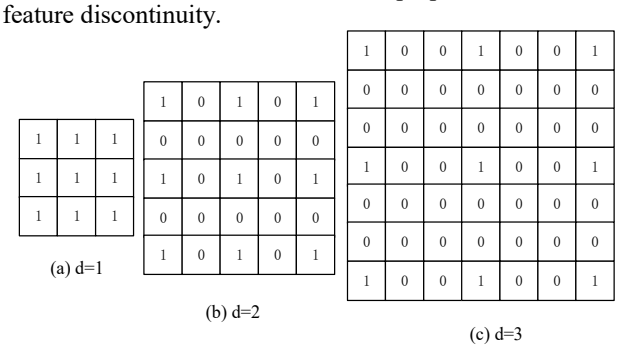


Figure 1. Dilated convolution

B. Cascade Residual Dilated Convolution

ResNet [11] uses a deeper network to extract more effective features through the idea of residuals. According

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00041

to the idea of residuals, in order to make effective use of the original features, this paper uses residual to fuse the input features and new features, which deepens the network structure and extracts more features. The fusion module as shown in Figure 2.

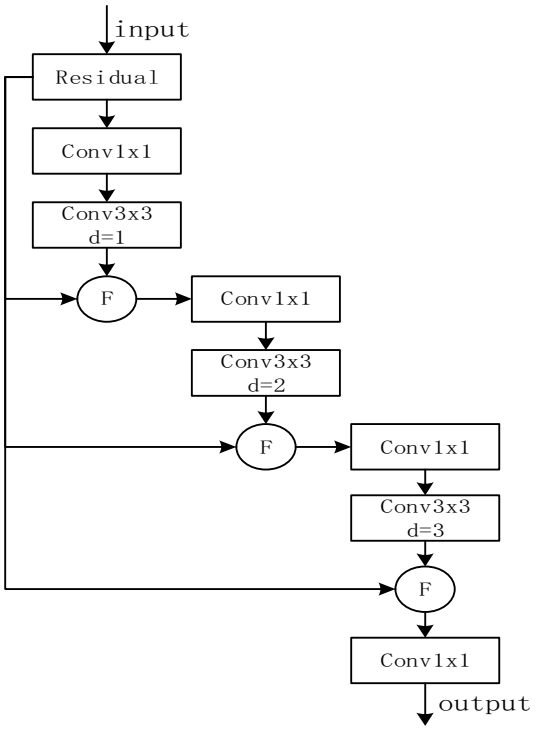


Figure 2. Cascade residual dilated fusion module

The dilated convolution with *d=1* is used to extract the local features of the small receptive field in the center of the convolution kernel. The dilated convolution with *d=2* is used to extract the surrounding features of the medium receptive field in the center of the convolution kernel. The dilated convolution with *d=3* is used to extract the long-distance features of large receptive fields in the center of the convolution kernel. In this paper, F represents add or concat for feature fusion.

C. Fusion Module with Coordinate Attention

In order to make the network pay attention to the features with accurate location information, Coordinate Attention(CA) [12] mechanism encodes the input features from the horizontal and vertical directions respectively according to formula 2 and formula 3, and decomposes the global pooling into the average pooling in two directions, so that the network can learn the location information of channel features and spatial features at the same time.

$$z_c^{h}(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h,i) \qquad (2)$$

$$z_c^{w}(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j,w) \qquad (3)$$

Based on the idea, when using cascaded dilated convolution to extract multi receptive field features, the target location information features are embedded into the hierarchical fusion module, and the network is combined with cascaded dilated convolution and location attention module to enhance the extraction of target features. A cascaded residual dilated fusion module embedded with location information is proposed as shown in Figure 3 below.

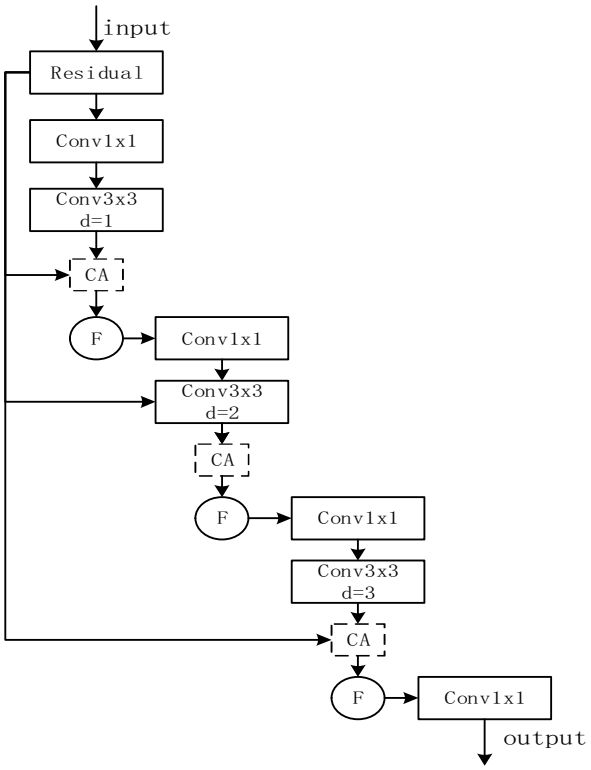


Figure 3. Cascaded residual dilated fusion module embedded with location information

III. OVERVIEW OF NETWORK STRUCTURE

Considering the detection speed, yolov4-tiny model is selected in this paper. Because it uses a shallow feature extraction network, it will lose some accuracy. In order to balance the detection accuracy and speed, the following improvements are made for the original yolov4-tiny:

(1) After the output feature of the second CSP(Cross Stage Partial) block, the proposed cascade residual dilated fusion module embedded with location information is added to improve the ability of network feature extraction by adding receptive fields;

(2) This module is also added to the second output feature of the backbone network, which aims to make the deep network pay more attention to spatial features through the module embedded with location information, combine semantic features with spatial features, and strengthen the learning of features.

So, the improved model dcca-yolov4-tiny detection structure of cascaded residual dilated fusion mechanism based on embedded location information is shown in Figure 4 below.

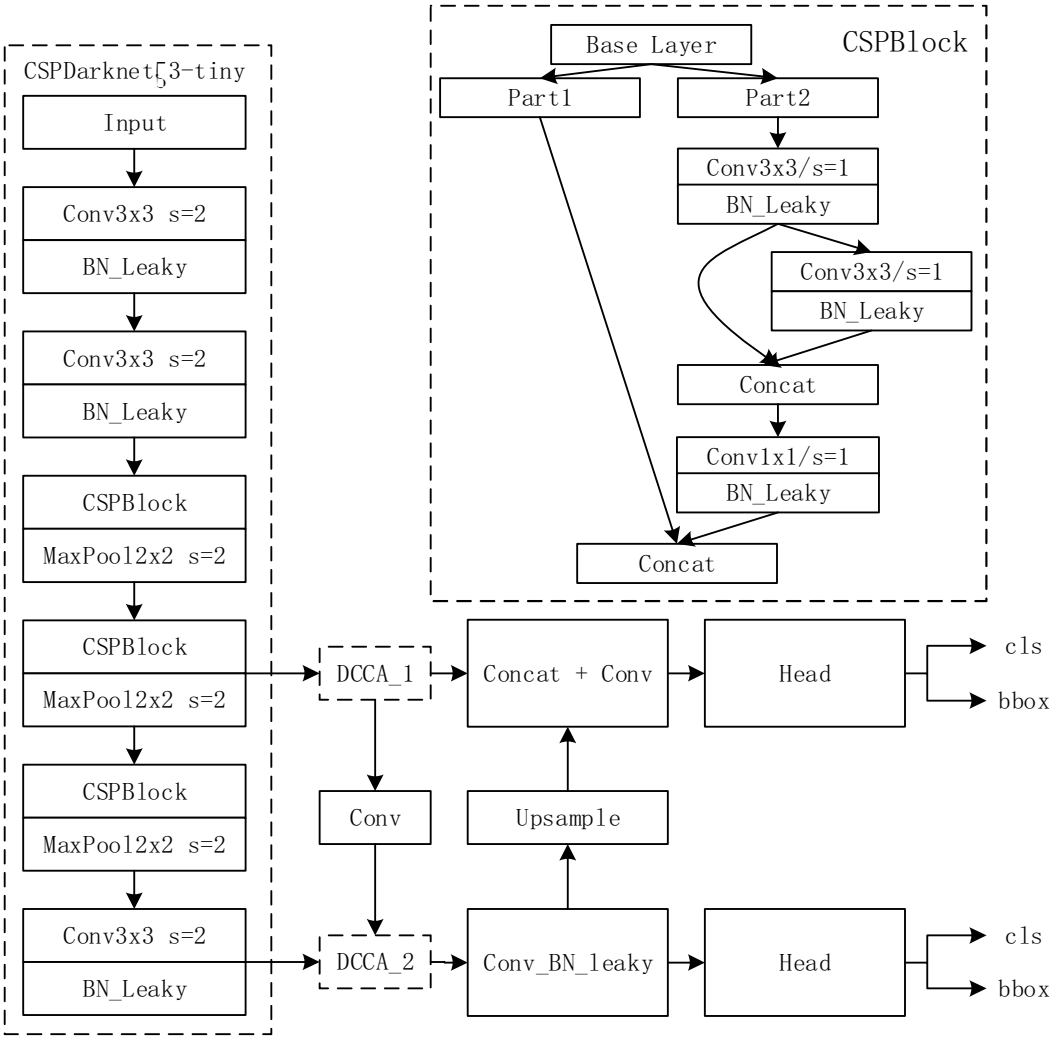


Figure 4. Structure of proposed dcca-yolov4-tiny

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this paper, experiments and evaluation are carried out on VOC data sets, including VOC2007(Visual Object Classes Challenge 2007) and VOC2012(Visual Object Classes Challenge 2012). This paper selects deep learning framework called PyTorch and uses 1080Ti GPU for training. The input picture size is 416x416, batch_size 32 is selected. This paper uses mAP(mean Average Precision) to verify the effectiveness of the model, and FPS(Frame Per Second) to evaluate the detection speed of the model.

Based on yolov4-tiny, this paper designs the following four modules: (1) the cascade residual dilated module using add fusion(dc_add); (2) the cascade residual dilated module using concat fusion(dc_concat); (3) the cascade residual dilated module embedded with location information using add fusion(dcca_add); (4) the cascade residual dilated module embedded with location information using concat fusion(dcca_concat). The results are shown in tables 1 below:

TABLE I. ORIGINAL YOLOV4-TINY AND OUR PROPOSED MODEL

	mAP(%)	FPS(frame/s)
yolov4-tiny	76.13	85
dc_add	78.00	74
dc_concat	79.03	69
dcca_add	79.12	72
dcca_concat	80.86	68

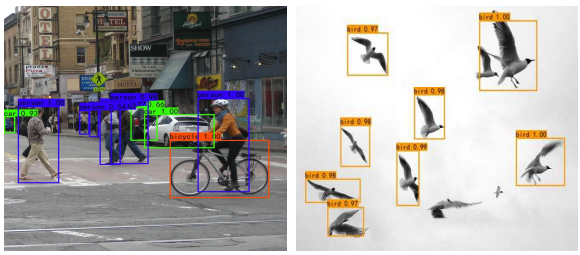
As shown in table 1, the mAP of dc_concat is 1.03% higher than that of dc_add, and the detection speed is reduced by 5FPS; the mAP of dcca_concat is 1.74% higher than that of dcca_add, and the detection speed is reduced by 4FPS. The experimental results show that the detection accuracy of the proposed module using different fusion methods is 1.87%, 2.90%, 2.99% and 4.73% higher than the original yolov4-tiny model, and the detection speed is lower than the original yolov4-tiny model, but it can still achieve the effect of real-time detection. Table 2 shows the detection results of each

category of the four different modules proposed in this paper.

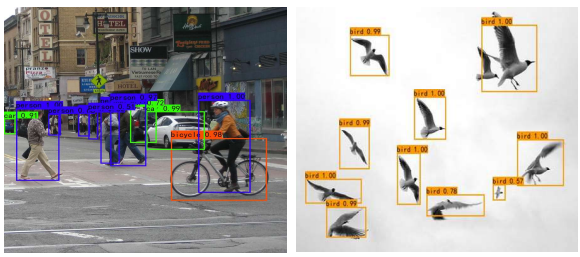
TABLE II. COMPARISON OF MAP OF THE PROPOSED MODULES

	dc_add	dcca_add	dc_concat	dcca_concat
car	91.43	**92.13**	92.21	91.97
horse	86.03	87.11	87.74	**89.25**
train	87.09	87.56	88.00	**89.02**
motorbike	87.65	85.30	86.04	**87.98**
person	86.05	85.72	85.82	**86.72**
bus	84.16	87.47	**88.70**	87.07
bicycle	85.86	85.49	85.94	**86.49**
cat	83.22	84.60	**86.00**	85.81
aeroplane	80.81	83.12	82.95	85.90
cow	82.47	**85.94**	85.86	85.89
sheep	81.28	83.15	**83.63**	83.31
tvmonitor	79.37	79.98	79.80	**81.57**
dog	76.62	79.54	79.48	**80.51**
bird	76.43	74.16	74.43	**79.21**
sofa	70.68	75.99	74.87	**78.30**
diningtable	71.94	75.38	72.65	**76.54**
boat	67.83	68.37	68.02	**71.32**
bottle	65.82	64.21	63.93	**68.43**
chair	62.54	62.98	61.50	**63.73**
pottedplant	52.90	54.22	53.02	**58.14**
mAP	**78.00**	**79.12**	**79.03**	**80.86**

In table 2, dcca_concat fusion method has the highest detection accuracy in small target boat, pottedplant and bird, which are 71.32%, 58.14% and 79.21% respectively, which is compared with dc_concat, the accuracy of small object detection can be improved by 3.30%, 5.12% and 4.78% respectively. It is proved that embedding location information into cascade residual fusion module can further improve the accuracy of small object detection.



(a) yolov4-tiny



(b) dcca-yolov4-tiny

Figure 5. Visualization of test results

Figure 5 above shows the visual results of the improved algorithm and the original yolov4-tiny model detection. It can be seen from the result on the left of the first picture that the original yolov4-tiny missed a person, but this paper uses the concat dense connection method in feature fusion, so the missed target is detected. Due to the use of dilated convolution to increase the receptive field and the way of embedding position information, it can be seen from the second picture that the improved dcca-yolov4-tniy model proposed in this paper detects the birds at the smallest lower right corner in the image.

V. CONCLUSIONS

Through the theoretical research on the attention mechanism of dilated convolution to increase receptive field and location information, a cascaded residual dilated fusion module embedded with location information is proposed in this paper, which is added to yolov4-tiny and used for object detection. The comparative experiments on VOC data sets show that the different modules proposed in this paper can improve the object detection accuracy of the original model. At the same time, the comparative experiments of feature fusion methods show that in practical application, add method can be selected from the perspective of parameters and model; In terms of feature reuse and accuracy, concat can be selected for fusion.

ACKNOWLEDGMENT

This work is supported by Hubei Provincial Key Laboratory of Transportation Internet of Things Technology under Grant 2017III028-002 and the Special Fund for Basic Scientific Research Business Expenses of Central Universities under Grant 2019III137CG.

REFERENCES

[1] Lin T, Dollar P, Girshick R, et al, Feature Pyramid Networks for Object Detection, Computer vision and pattern recognition,2017.

[2] Tan M, Pang R, Le Q V J a P A, Efficientdet:Scalable and efficient object detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[3] Liu S, Qi L, Qin H, et al, Path Aggregation Network for Instance Segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[4] LIU S, HUANG D, Receptive field block net for accurate and fast object detection, Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[5] FU C Y, LIU W, RANGA A, et al, DSSD:deconvolutional single shot detector, https://arxiv.org/abs/1701.06659.

[6] XU X F, ZHAO W F, ZOU H Q, et al, Detection algorithm of safety helmet wear based on MobileNet-SSD, Computer Engineering, vol. 47, pp. 298-305, 2021.

[7] GAO Q Q, HUANG B C, LIU W Z, et al, Detection method of bamboo strip surface defects based on improved CenterNet, Journal of Computer Applications, vol. 41, pp. 1933-1938, 2021.

[8] REDMON J, DIVVALA S, GIRSHICK R, et al, You only look once:unified, realtime object detection, IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[9] Bochkovskiy A, Wang C Y, Liao H Y M, YOLOv4: Optimal Speed and Accuracy of Object Detection, https://arxiv.org/abs/2004.10934.

[10] YU F, KOLTUN V, FUNKHOUSER T, Dilated residual networks, Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[11] HE K M, Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2016.

[12] HOU Q B, ZHOU D Q, FENG J S, Coordinate Attention for Efficient Mobile Network Design, https://arxiv.org/abs/2103.02907.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

An effective lightweight attention mechanism

1st Zhipeng Liu
School of Artificial Intelligence and Computer Science
Jiangnan University
Wuxi, China
952738565@qq.com

2nd Wei Fang
School of Artificial Intelligence and Computer Science
Jiangnan University
Wuxi, China
fangwei@jiangnan.edu.cn

3rd Jun Sun
School of Artificial Intelligence and Computer Science
Jiangnan University
Wuxi, China
junsun@jiangnan.edu.cn

***Abstract*—Aiming at the problem of large parameters and poor portability of attention mechanism modules, an extremely lightweight attention mechanism is designed, which uses a combination of spatial attention and channel attention to enhance the model's attention on important information. From the experimental results, the lightweight attention mechanism can add to the deep convolutional neural network which introduces with negligible parameters and calculations, allows the model to focus more on useful information and improve the accuracy of the object detection task. It surpasses SENet(Squeeze and Excitation Networks) in performance and is easily transplanted to mainstream deep convolutional neural networks such as ResNet.**

Keywords-Convolutional neural network; Object detection; Attention mechanism; Lightweight;

I. INTRODUCTION

In the context of the era of big data and artificial intelligence, the rapid development of deep learning has led to major developments in computer vision tasks. Object detection has also developed vigorously as a core task in the field of computer vision. However, due to the large scale of data, the complexity and variety of data sources, ever-changing environmental backgrounds, and limited computing resources, target detection tasks still face many challenges, and they are always struggling on the road to improve detection accuracy and detection speed.

Object detection [1] is one of the important tasks of computer vision, and it is considered to be the main pioneer in solving the problem of semantic understanding of the surrounding environment. Object detection is the basis for solving complex high-level vision tasks such as segmentation, scene understanding, target tracking, image description, event detection and activity recognition. Object detection has a wide range of applications in many fields of artificial intelligence and information technology, including robot vision, autonomous driving, human-computer interaction, content-based image retrieval, intelligent video surveillance, and augmented reality. In human vision, the attention mechanism is inseparable at all times, as a result of it exists all the time, it is often ignored. For example, when reading books and newspapers, people will pay attention to pictures and text information. At this time, information such as paper color and material will not be the focus. Humans will treat different information differently and pay more attention to valuable information.

Nowadays, deep learning [2] has developed rapidly. Computer vision is mostly based on neural networks [3]. Neural networks can self-learn and extract features from input pictures and other information. For these features, we cannot generalize and perform equivalent processing, because different features play different roles for reasoning. The degree of help is different, so it is particularly important for the neural network to construct the attention mechanism. Realize the neural network to focus on the effective area autonomously, reduce the attention to irrelevant information, and improve the discriminative ability of the network, which is the core of the attention mechanism in computer vision. The neural network can learn from itself, which provides the possibility for parameter learning of the attention structure. At the same time, the attention mechanism can allow humans to better understand the world expressed by the neural network, and provide a visual basis for network structure optimization and parameter adjustment.

In recent years, most of the research on the attention mechanism of computer vision is based on masks to generate attention structure modules. In the training process of neural networks, the required attention areas in the pictures are learned, the weights are updated, and the masks are generated. It is weighted with the original features to form an attention mechanism, which in turn allows the network to pay more attention to the features of discriminative regions. Mask-based attention mechanisms are divided into soft attention mechanisms and hard attention mechanisms. The soft attention mechanism pays more attention to the channel or spatial information, which can be differentiated. This allows the parameters of the soft attention structure to be calculated in the process of backpropagation of the neural network, learn and adjust during the iterative process. After the training is completed , soft attention can be directly generated by the network, so soft attention is also called determinable attention. Hard attention pays more attention to pixels, emphasizing the attention of each pixel, similar to a binary method, its prediction is a dynamic random process, and it is a kind of non-differentiable attention that cannot be self-learned through the network. Realization is usually done through reinforcement learning.

Since soft attention can realize self-learning in neural networks and is more widely used, this paper mainly studies the soft attention mechanism and proposes a lightweight attention mechanism that is easy to embed in existing neural networks to improve detection tasks effect.

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00042

II. METHODS

From the perspective of spatial attention and channel attention merged, a lightweight attention mechanism is proposed. The overall structure of the attention mechanism is shown in Figure 1. For the input feature map, the mask is obtained by the spatial attention module and then multiplied and weighted by the original feature map. The output result of the spatial attention module and the mask generated by the channel attention module are multiplied and weighted to obtain the final output. The output feature map has the same dimension as the original input feature map. Representatives of spatial attention mechanism are STN [4], SMCA [5], and representative of channel attention mechanisms are SENet [6], FcaNet [7].

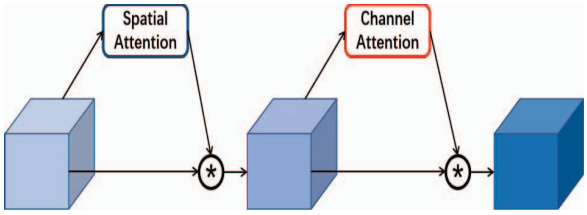


Figure 1. The Architecture of Designed Attention Mechanism.

A. Design of Spatial Attention Module

The purpose of the spatial attention module is to judge the importance of spatial pixels. The structure is shown in Figure 2. For the input feature map, the average value of all channels at each pixel position is calculated as the initial value of the importance for the pixel. Then, the 3×3 convolution and the dilated convolution [8] with the expansion rate of 3 are used in parallel. This operation uses different receptive fields to fully perceive the importance of a pixel compared to the surrounding space pixels, and the use of dilated convolution can increase the receptive field, without introducing additional parameters, the results obtained by the parallel convolution branch are connected to the channels, the feature importance is merged through 1×1 convolution, and then the spatial attention mask and the meta input feature are obtained through the sigmoid function The graph is multiplied and weighted to get the final output.

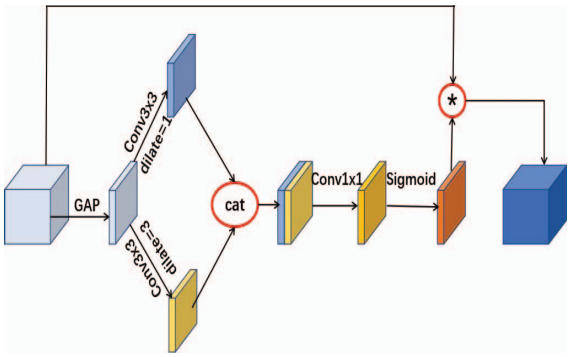


Figure 2. The Architecture of Spatial Attention Module.

B. Design of Channel Attention Module

The purpose of the channel attention module is to judge the importance of channel features. The structure is shown in Figure 3. For the input feature map, global pooling is performed to obtain the incident channel information, and then the channel attention is obtained through a parameterized weighting method. This parameterization method introduces very few parameters, which ensures the lightweight characteristics of the attention mechanism. Finally, the channel attention mask is weighted and multiplied by the original input feature map to obtain the final output.

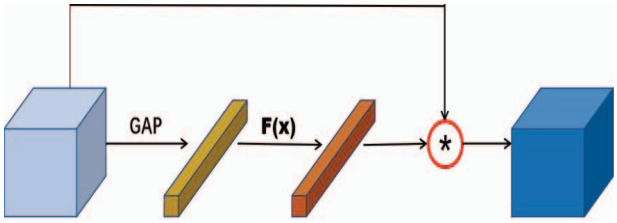


Figure 3. The Architecture of Channel Attention Module.

III. EXPERIMENT

A. Experimental Setup

The experiment is based on python 3.6, uses the pytorch framework to build the network, and is carried out in the Linux system environment, uses the GPU model Tesla K80. The experiment is based on the SSD [9] algorithm on the VOC [10] data set. The input image size is 300×300, and the designed attention mechanism is embedded in ResNet18, ResNet34, ResNet50 and ResNet101. By this way, we can obtain DA-ResNet18, DA-ResNet34, DA-ResNet50 and DA-ResNet101, the structure of the DA module embedded in ResNet is shown in Figure 4.

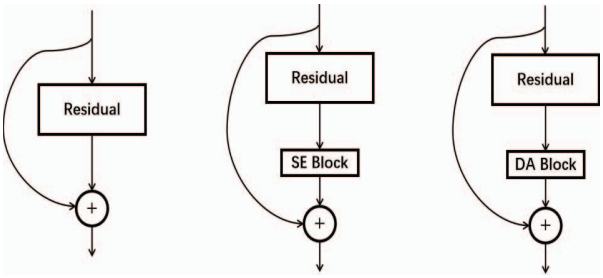


Figure 4. The architecture of DA embedded residual network

B. Evaluation Standard

FLOPs (Floating Point Operations) represent the number of floating point operations, which are usually used as indicators to measure the complexity of the model. For the convolutional layer, the formula of FLOPs is shown as follows. W and H respectively represent the width and height of the input feature vector, C_{in} represents the number

of channels of the input vector, K represents the size of the convolution kernel, and C_{out} represents the number of channels of the output feature vector.

$$FLOPs\_conv = 2 \times H \times W\left(C_{in}K^2 + 1\right)C_{out} \quad (1)$$

For the fully connection layer, the formula of FLOPs is calculated as follows. I represents the dimension of the input and O represents the dimension of the output.

$$FLOPs\_fully = \left(2I - 1\right)O \quad (2)$$

Parameters are also an important indicator to measure the complexity of the model. The formula of calculation is shown as follows. The meaning of the variable is the same as before.

$$Params = C_{out} \times \left(H \times W \times C_{in} + 1\right) \quad (3)$$

Grad-CAM [11] is a gradient-based visualization algorithm that can differentiate different areas in the picture by generating a heat map, so that we can intuitively understand the area of a picture that the neural network cared about, and it is useful for optimizing experiments and understanding the neural network.

Mean Average Precision is used as the evaluation standard of algorithm accuracy, we usually called it mAP. The calculation of mAP is closely related to IoU. The formula of IoU is shown as follows. b_{gt} represents the position information of the real box and b_{pred} represents the position information of the predicted box. By this, we get all boxes' IoU.

$$IoU\left(b_{pred}, b_{gt}\right) = \frac{Area\left(b_{pred} \cap b_{gt}\right)}{Area\left(b_{pred} \cup b_{gt}\right)} \quad (4)$$

The IoU of the predicted box and the real box is calculated, and then the default box is divided into positive or negative examples according to the set IoU threshold. In the image, each category will have real cases (TP, True Positive), false positives (FP, False Positive), true negatives (TN, True Negative), and false negatives (FN, False Negative). The formula of precision is shown as follows.

$$precision = \frac{TP}{TP + FP} \quad (5)$$

Choose 11 different recall rates (0, 0.1, 0.2,..., 0.9, 1.0), calculate the corresponding accuracy rates under different recall rates, and then average the accuracy rates to get the AP, mAP is the average value of all types of AP. The formula of AP is shown as follows.

$$AP_{11point} = \frac{1}{11}\left(\sum x\right)\left(x \in \max precision\right) \quad (6)$$

C. Experimental Result

In order to explore the best combination of the spatial attention module and the channel attention module, the serial and parallel methods of the two method are tested. The experimental results are shown in Table 1. It is not difficult to see from the experimental results that the channel attention is in the front and the spatial attention is in the back, the spatial attention is in the front and the channel attention is in the back, and the parallel combination of the two module can all improve the detection accuracy. The serial mode which the spatial attention is in the front and the channel attention is in the back achieves the highest accuracy. As a result, the combination of the spatial attention part and the channel attention part of the designed lightweight attention module is determined.

TABLE I. Accuracy results of different combinations of attention modules

Method	mAP
ResNet50	75.9
ResNet50+channel+spatial	76.8
ResNet50+spatial+channel	**77.0**
ResNet50+spatial and channel in parallel	76.6

Compare the parameters and calculations of ResNet18, ResNet34, ResNet50, ResNet101, SE-ResNet18, SE-ResNet34, SE-ResNet50, SE-ResNet101, DA-ResNet18, DA-ResNet34, DA-ResNet50, DA-ResNet101, and use the above-mentioned network as the backbone feature extraction network. SSD is used as the baseline to compare the accuracy. The experimental results are shown in Table 2. From the experimental results, we can see that SE-ResNet50 has increased 2.5150M and 0.0026GFLOPs compared to ResNet50 in terms of parameters and calculations, while DA-ResNet50 has only increased by 0.0004M and 0.0003GFLOPs compared to ResNet50 in terms of parameters and calculations. The amount of our DA attention mechanism's parameters and calculations are negligible. What's more, similar conclusions can be obtained from the comparison results of other items in the table, and our method's accuracy of the detection task is improved more significantly than that of SENet. The results improve that our method is very effective.

TABLE II. DA attention mechanism vs. SE attention mechanism

Backbone	Params	GFLOPs	mAP
ResNet18	11.6895M	1.8214	72.6
SE-ResNet18	11.7766M	1.8215	72.9
DA-ResNet18(ours)	**11.6897M**	**1.8215**	**73.1**
ResNet34	21.7977M	3.6742	74.5
SE-ResNet34	21.9549M	3.6744	74.8
DA-ResNet34(ours)	**21.7980M**	**3.6744**	**75.1**
ResNet50	25.5570M	4.1185	75.9
SE-ResNet50	28.0720M	4.1211	76.5
DA-ResNet50(ours)	**25.5574M**	**4.1188**	**77.0**
ResNet101	44.5492M	7.8444	77.0
SE-ResNet101	49.2923M	7.8492	77.6
DA-ResNet101(ours)	**44.5498M**	**7.8448**	**78.2**

The experiment compared the parameters, calculations and accuracy of the DA module with other lightweight attention mechanisms on the VOC detection task. As shown in Table 3, it can be seen from the experimental results that the amount of parameters and calculations of DA module we

designed is less compared with SGE [12] and ECA [13] attention modules, the accuracy is higher.

TABLE III. DA attention mechanism vs. other attention mechanisms

Method	Params	GFLOPs	mAP
ResNet50	25.557M	4.119	75.9
+SGE	25.557M	4.127	76.3
+ECA	25.559M	4.127	76.8
+DA(ours)	**25.557M**	**4.119**	**77.0**
ResNet101	44.549M	7.844	77.0
+SGE	44.553M	7.858	77.4
+ECA	44.549M	7.858	78.0
+DA(ours)	**44.550M**	**7.845**	**78.2**

Here, ResNet50 and DA-ResNet50 are visualized by using Grad-CAM in layer4_2, as shown in Figure 5. In the same row of pictures, the leftmost is the original picture, the middle is the result of ResNet50 visualization, and the far right is the result of DA-ResNet50 visualization. It is not difficult to see that when the DA module is added, the network pays more attention to effective information, and the area of interest is also more concentration, which proved the effectiveness of the DA attention module.

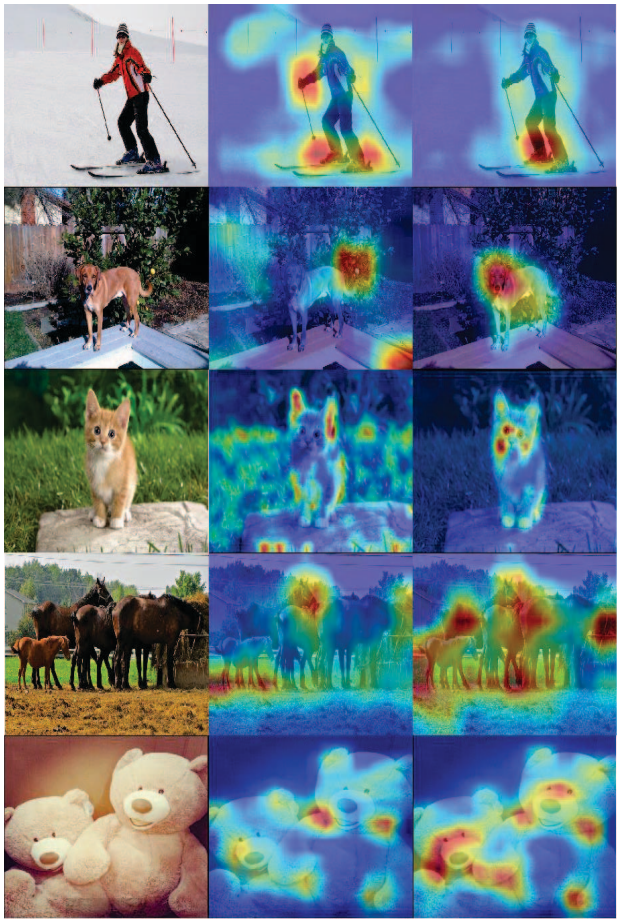


Figure 5. Grad Cam Visualization Comparison

IV. CONCLUSION

Aiming at the problem of large parameters and poor portability of attention mechanism modules, an extremely lightweight attention mechanism is designed, which uses a combination of spatial attention and channel attention to enhance the model's attention to important information. The attention mechanism we proposed is extremely lightweight, which introduces very few parameters and calculations, and improves the detection accuracy significantly. It is easy to be embedded in deep convolutional neural networks such as ResNet, and has strong portability. Experiments have proved that our attention mechanism is very efficient.

The attention mechanism designed in this paper has fewer parameters and calculations, but the improvement effect of the model needs to be further improved. How to design an attention mechanism with fewer parameters, strong portability, less calculation, and obvious improvement effect is the direction of our future work.

ACKNOWLEDGMENT

We wish to thank every member of the team for their efforts. Thank you for your help in my study. I will continue to be enthusiastic in the field of object detection and conduct in-depth research.

REFERENCES

[1] Wu X, Sahoo D, Hoi S. Recent Advances in Deep Learning for Object Detection[J]. Neurocomputing, 2020, 396.

[2] Zhang Q, Yang L T, Chen Z, et al. A survey on deep learning for big data[J]. Information Fusion, 2018, 42:146-157.

[3] Ketkar N. Convolutional Neural Networks[J]. Springer International Publishing, 2017.

[4] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[J]. arXiv preprint arXiv:1506.02025, 2015.

[5] Gao P, Zheng M, Wang X, et al. Fast Convergence of DETR with Spatially Modulated Co-Attention[J]. arXiv preprint arXiv:2101.07448, 2021.

[6] Hu J, Shen L, Sun G.Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[7] Qin Z, Zhang P, Wu F, et al. FcaNet: Frequency Channel Attention Networks[J]. arXiv preprint arXiv:2012.11879, 2020.

[8] Liu S, Huang D. Receptive field block net for accurate and fast object detection[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 385-400.

[9] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[10] Vicente S, Carreira J, Agapito L, et al. Reconstructing pascal voc[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 41-48.

[11] Selvaraju R R , Cogswell M , Das A , et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization[C]// IEEE International Conference on Computer Vision. IEEE, 2017.

[12] Li X , Hu X , J Yang. Spatial Group-wise Enhance: Improving Semantic Feature Learning in Convolutional Networks[J]. 2019.

[13] Wang Q , Wu B , Zhu P , et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[J]. 2019.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Traceability method between design documents and source codes based on SQL dependency

Lujing Yu,Yonghua Li
School of Computer Science and Technology,
Wuhan University of Technology
Wuhan, China
Email: 1591891342@qq.com, liyonghua@whut.edu.cn

Yuqing Feng
Academic Affairs Office,
Wuhan Social Work Polytechnic
Wuhan, China
Email: 280064139@qq.com

Chen Qi
Information Technology Department,
Chongqing Guoyuan Port Co, Ltd
Chongqing, China
Email: 358793710@qq.com

***Abstract*—Information Retrieval (IR) technology was widely used in traceability between design documents and source codes. However, the vocabulary mismatch between the design documents and the source codes affects the performance of IR. Aiming at the above situation, a dynamic tracing method from design documents to source codes combining IR technology and SQL statement is proposed in management information system. Firstly, the similarity of the two is calculated by IR and the candidate links are generated; Then, the SQL statement required by the codes is automatically estimated according to the design documents, and the SQL statement is compared with the actual SQL statement in the codes to correct the design documents-codes similarity score; Finally, set a threshold to determine the trace links of the design documents to the source codes. The experimental results show that this method can improve the similarity score of code classes with relevant SQL statements in the design documents, so as to improve the ranking of code classes in the candidate links, extract the trace links that may be missing in IR method under the action of threshold, and finally improve the precision of trace results.**

***Keywords*-software traceability; SQL dependency; dynamic traceability; information retrieval; software engineering;**

I. INTRODUCTION

Software Traceability refers to the ability to interrelate any uniquely identifiable software engineering artifact to any other, maintain required links over time, and use the resulting network to answer questions of both the software product and its development process[1]. There is a correlation between documents and codes. Discovering and maintaining the tracking links between design documents and codes helps software engineering activities such as program understanding, software maintenance, and requirements tracking.

At present, the most widely used software tracking method is Information Retrieval (IR), that is, by constructing IR model and identifying the tracking links according to the text similarity between software artifacts. Literature[2] first used IR technology to generate the tracking links of software documents and program codes. On the theoretical basis of early work[3], literature[4] proposes a concept of code dependency, called Closeness, to quantify the degree of interaction between classes based on code dependency and data dependency. Literature[5] extends the previous work combining direct code dependency and IR technology[6] through User Feedback. Literature[7] proposed a demand tracking method combining IR technology and non-text technology.

Most of the above researches firstly calculate the similarity from source codes to target file based on IR similarity model, establish the list of candidate links, then combine with other technologies (such as code dependency) to weight the similarity of candidate links, and finally set the threshold get the final tracking link. However, due to the different vocabulary usage between the documents and the codes, it is difficult to match, so these methods based purely on IR model are usually less accurate in establishing the documents-codes trace link.

This article is aimed at information management systems. In such systems, there are usually close database interactions. There are bound to be database operation statements in the design documents. These statements will be related to the SQL statements in the source codes[8]. Based on the description information in the design document, this paper predicts to generate the corresponding SQL statements, analyzes the correlation between the predicted SQL statements and the SQL statements in the actual codes, and optimizes the trace links generated by IR technology according to the analysis results.

II. TRACING METHOD BASED ON SQL DEOENDENCY

The process of designing documents to codes tracing method based on SQL dependency is shown in Figure 1. It consists of five processing stages: first, the material is pretreated to obtain the domain term material; Then IR model is used to generate candidate links list between source material and code classes, generate source material (estimated SQL statements) evaluation unit token sequence, and generate actual SQL statements evaluation unit token sequence. Finally, the SQL dependency closeness between the source material evaluation unit and the actual SQL evaluation unit is calculated and analyzed.

A. Material preprocessing

This paper extracts the title and function description from the design document as the source material docu-

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00043

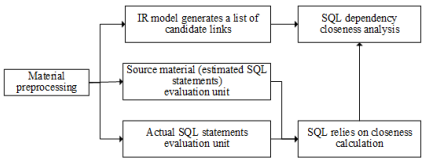


Figure 1. processing flow based on the SQL dependency tracking method.

ment; from the source code document, extracts the method comment, method body, and method namespace as the target material document. The actual SQL statements are also extracted from the source code document. Finally, this step extracts the database table information and the relationship between the tables from the database physical data model (PDM) document, as the domain term material.

B. IR model generates a list of candidate links

This paper uses the Vector Space Model (VSM) method[2] to generate a list of candidate links between the source material and the code classes, including the following 4 steps: 1) create a Corpus; 2) Normalized the corpus; 3) Make corpus and calculate the text similarity; 4) Generate a list of candidate links

C. Cgeneration of evaluation unit sequence of source material(estimated SQL statement)

The generation of source material evaluation unit sequence needs to parse the PDM document and the establish Database Table Information Struct (DTIS), and then extract the domain-related terms from the design document to establish a description term mapping set, which is used to map SQL operation type and SQL Predicate, etc. Before the conversion, Stanford Parser[9] is used to analyze the source material to determine the phrase corresponding to the phrase in the sentence and the part of speech of the word.

D. Generation of actual SQL statement evaluation unit sequence

The accuracy of the trace links can be seriously affected by the possible differences in variable names and structures between actual and estimated SQL statements. Therefore, before the automatic tracing, the actual SQL statement needs to be normalized while ensuring the semantics. The conversion process of the evaluation unit sequence includes two steps: normalization of SQL sentences and lexical analysis.

E. SQL dependency and SQL dependency closeness calculation method

1) SQL dependency: This paper refers to the relationship between the estimated SQL statements of the source material and the actual SQL statements as SQL dependency. The estimated SQL statements mentioned in this article refer to the SQL object converted from the source material combined with DTIS and the description term mapping set, for example, SQL keywords such as SELECT,WHERE, and database table or field names.

2) SQL dependency closeness calculation method: SQL Dependency closeness: Measure the correlation between the SQL object converted from the source material and the dependency of actual SQL statements.

a)Search item setting

This article compares the estimated SQL statements with the actual SQL statements from the following six aspects:

- F_keyword: Refers to keywords used to describe the structure of SQL.
- F_table: SQL statement operation master data table.
- S_table: F_table is a table related by a foreign key, also called a slave table.
- F_field: Represents the operation field.
- O_field :Represents the condition field.
- O_keyword: Other keywords (such as FROM, WHERE, etc).

Assuming that the weight of each unit type is W_i (i = 1-6) and the weight ratio of SQL at different levels is $(1/2)^{level-1}$, the token sequence of estimated SQL statements is X, and the token sequence of actual SQL statement is Y. The similarity calculation formula of sequence X and Y is shown in equation (1) :

$$sim(X,Y) = \begin{cases} \sum_{i=1}^{level-1} \frac{1}{2} \sum_{j=1}^{6} \frac{n_j}{N_j} \times W_j \\ 0, The\ F\_keyword\ of\ X \neq \\ \quad F\_keyword\ of\ Y \end{cases} \quad (1)$$

Where, n_j represents the number of matches of each unit type in the Token sequence of X and Y, N_j represents the number of each unit type in Y, W_i represents the corresponding weight, and $level$ represents the number of levels of SQL statements.

b) The steps to calculate the similarity of SQL statements

step 1: Analyze the source material and convert it into SQL object to form evaluation unit sequence (token sequence) X.

step 2: Tokenize the actual SQL statements in the source program, divide the evaluation units according to the order of the SQL statements and the separator specified by the syntax to form a token sequence Y, and count the number of types of units in Y is N_j.

step 3: Count the matching number n_j of each unit type between X and Y sequence at the same level; at the same time, calculate the similarity of SQL at a certain level by equation (1).

step 4: If the level of the actual SQL statement is $level$¿1, repeat step 2 to obtain the overall similarity of the X and Y sequences.

F. SQL dependency closeness analysis

This section will introduce the IR value weighting algorithm between the source material and the code classes, so as to improve the IR value of candidate links generated in Section 1.2 and realize the reordering of the candidate link list. The weighted algorithm can not only weight the IR value of the candidate links of the class in which the actual SQL statement is located, but also get corresponding

weights for other classes that have a call relationship with the class.

Definition1 : The token sequence and the weight of each retrieval item are:

$$W_{token} = (W_{F\_keyword}, W_{F\_table}, W_{S\_table}, W_{F\_field}, W_{O\_field}, W_{O\_field})$$

The calculation equation of the score points is shown in equation (2).

$$W_{token} = \begin{cases} W_{F\_keyword} = 0 \\ W_{F\_table} = W_{init} \\ W_{S\_table} = 1/2 W_{W_F\_table} \\ W_{field} = 1/4 W_{F\_table} \\ W_{O\_field} = 1/20 W_{F\_table} \\ W_{O\_keyword} = 1/20 W_{F\_table} \end{cases} \quad (2)$$

Through the test of the data set, it is found that when W_{init}=0.2, the overall effect of improving the candidate links list based on SQL dependency closeness is the best, so 0.2 is selected as W_{init}.

Definition2 : Suppose the target class set of the candidate links list is C_T, the target class C_{Ti} calls the C_{DAL}-like method set in the DAL layer as M, and the weight that C_{Ti} can obtain based on the SQL dependency is W_{SQL}. The weight of C_{DAL} is W_{DAL}.

The calculation equations of W_{SQL} and W_{SQL} are shown in equation (3) and equation (4):

$$W_{SQL} = \frac{1}{n}(\sum_{i=0}^{N} sim(X, Y_i)) \quad (3)$$

$$W_{DAL} = Bonus_{max} = Max(W_{SQLi}), i = 0, 1, ..., |C_T| \quad (4)$$

Among them, N is the number of methods $|M|$, and Y_i represents the token sequence of the SQL statements in the method $m_i (m_i \in M)$. Based on the SQL dependency closeness analysis, weight the IR values of C_{SQL} and C_{DAL}.

III. Experiments and analysis

A. Experimental data set

The experimental materials in this paper are from the design documents (source material) and sources code (target material) of a port production business management system (PCTOS) in Chongqing. The PCTOS includes four subsystems: business, dispatch, tally, and materials. Table 1 records the source material and target material types, the number of materials, and the number of correct tracking links that need to establish a tracking link relationship in the PCTOS system. In this paper, Visual Studio 2015 was used to export the code diagram DGML file and obtain the calling relationship of the classes. At the same time, the PDM document of PCTOS is also used in the experiment to obtain the domain corpus.

B. Analysis of results

For the setting of the threshold, this experiment first tests the material system data set, and uses the F2Measure metric to evaluate the accuracy of the tracking link returned on the 4 sets of data (averaged). F(0.1)

Table I
Experimental data set, materials and number of correct links

Subsystem	Number of materials (pieces)		Correct links (pieces)
	Source material	Source material	
Business	32	132	96
Dispatch	30	120	92
Tally	26	110	81
Meterials	32	140	101

represents the threshold $Thresholdlink$=0.1, similarly $Thresholdlink$=0.2, 0.3 corresponds to F(0.2), F(0.3). When Thresholdlink¿0.3, several comparison methods, tracking link recall is greatly reduced, while the accuracy is not significantly improved, so this paper selects 0.1, 0.2 and 0.3 to set the threshold.

From Table 2, the F(0.1) of the four sets of data based on the SQL dependency closeness analysis method is higher than F(0.2), F(0.3), so the threshold $Thresholdlink$=0.1 is the most appropriate; although the VSM method is in the first F(0.2) is the largest in the two sets of data (16 samples), but F(0.1) is the highest in the other sets of data, so $Thresholdlink$=0.1; in the same way in literature[7], threshold $Thresholdlink$=0.2.

The box plot generated by the R tool[10] is shown in Figure 5. The X axis represents the three methods compared in the article, and the Y axis represents the recall, precision, and F2Measure of the three methods. The box plot shows the minimum, the first quartile (the lower line of the box), the median (the line in the middle of the box), the third quartile (the upper line of the box), and the maximum from bottom to top. Any point that falls further will be considered an "extreme" value and will be drawn separately. It can be seen from the results in Figure 2 that the recall rate based on the SQL dependency closeness analysis method and the literature[7] method is not obvious compared to the VSM method, but the precision rate and F2Measure are better. In terms of precision and F2Measure indicators, the method in this paper is superior to the other two methods in terms of median and first quartile. This shows that the analysis method of combining IR technology and SQL dependency proposed in this paper is effective for the correction of the candidate links list generated by VSM.

IV. Conclusion

This paper proposes a method of extracting trace link from design documents to source codes based on SQL dependency. The method first extracts the function description from the design document with chapter granularity as the source material, extracts key information such as method content and method namespace from the source code document with the method granularity as the target material, and uses the VSM model to calculate the similarity score between the source material and the target material, and then obtain the candidate links list of the source material and the code, sorted in reverse order by

Table II
F2MEASURE INDEX EVALUATION RESULTS

Data Group	VSM			Literature[7]			Method of this article		
	F(0,1)	F(0,2)	F(0,3)	F(0,1)	F(0,2)	F(0,3)	F(0,1)	F(0,2)	F(0,3)
8	0.565	0.532	0.524	0.616	0.601	0.551	0.685	0.643	0.589
16	0.534	0.567	0.507	0.563	0.592	0.521	0.642	0.630	0.605
24	0.570	0.487	0.461	0.539	0.587	0.526	0.658	0.601	0.587
32	0.572	0.530	0.473	0.548	0.593	0.515	0.690	0.598	0.563

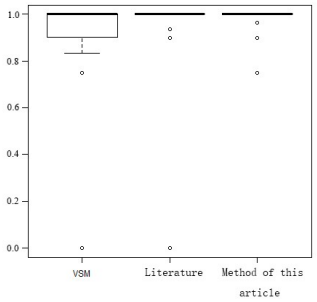


Figure 2. Recall

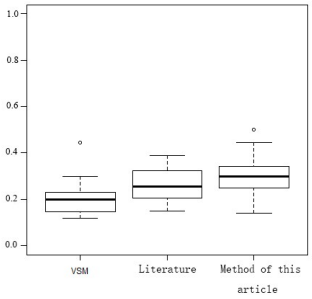


Figure 3. Precision

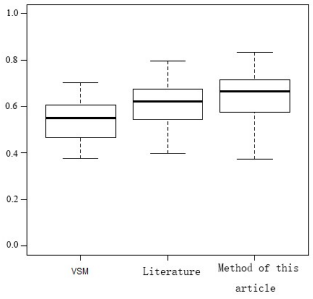


Figure 4. F2Measure

Figure 5. Box plot of experimental results.

the IR value; then combine the dependency relationship between the source material and the actual SQL statements in the code, calculate the SQL dependency closeness of the source material and the actual SQL statements, and then obtain the weight of the source material and the code category, and adjust the IR value in the source material and the code category candidate links; Finally, the threshold is set to determine the tracking link of the source material and the code category. For the same experimental object, this method improves the precision rate compared with the traditional IR method. However, when converting the source material into the evaluation unit sequence, only the single-level SQL structure is considered at this stage, and the nesting is not involved. Second, the set of rules used to eliminate ambiguity in the conversion process is limited, which can be continuously expanded in subsequent work. In the follow-up work, user feedback will also be combined with the method in this article to achieve better results.

REFERENCES

[1] CoEST. Center of excellence for software traceability[EB/OL]. [2020-06-02]. http://www.CoEST.org.

[2] Salton G,Wong A,Yang C S. A vector space model for automatic indexing[J].Communications of the ACM,1975, 18(11): 613-620.

[3] Mona Rahimi ; Jane Cleland-Huan. Evolving Software Trace Links between Requirements and Source Code[J]. Empir. Softw. Eng, 2018,23(4):2198-2231.

[4] Hongyu Kuang, Hui Gao, et al.Using frugal user feedback with closeness analysis on code to improve IR-based traceability recovery[J]. ICPC 2019: 369-379.

[5] Panichella A, Mcmillan C, Moritz E, et al.When and How Using Structural Information to Improve IR-Based Traceability Recovery[J]. The 17th European Conference on Software Maintenance and Reengineering, 2013, 88(2):199-208.

[6] Shiheng Wang , Tong Li , Zhen Yang. Exploring Semantics of Software Artifacts to Improve Requirements Traceability Recovery: A Hybrid Approach[C]. Asia-Pacific Software Engineering Conference (APSEC),2019.

[7] JYOTI,CHHABRA J K.Requirements traceability through information retrieval using dynamic integration of structural and co-change coupling[C]// Proceedings of the International Conference on Advanced Informatics for Computing Research.Berlin German: Springer,2017: 107-118.

[8] Hang Li. Learning to Rank for Information Retrieval and Natural Language Processing: Second Edition [M]. Morgan & Claypool,2014.

[9] Stanford University. The Stanford Parser: A statistical parser[EB/OL] [2017-06-09]. https://nlp.stanford.edu/software/lex-parser.shtml

[10] Auckland University.R[EB/OL].(2020-06-20),[2020-07-30].https://www.r-project.org.

KTSP Solution Based on Workload Balance and Genetic

Zejun Hu
College of artificial intelligence
Guangdong mechanical & electrical polytechnic
Guangzhou, Guangdong 510515,China
Email:huzejun@sina.com

Guohua Yang
College of artificial intelligence
Guangdong mechanical & electrical polytechnic
Guangzhou, Guangdong 510515,China
Email: 41759307@qq.com

***Abstract*—This paper focuses on the KTSP (K-person Traveling Salesman Problem) programming problem of workload balance. By adding a virtual central node, the KTSP problem is described as a TSP(Travelling Salesman Problem) problem with workload balance constraints. Then, 3-opt(3-optimization) is used for local optimization and genetic method is used to search the global optimal solution. Finally, the qualified solution is limited by workload balance constraints, and the workload balance algorithm is used to correct the results. The planning method in this paper is effective for the medium-sized workload balancing KTSP problem, and can obtain the ideal solution.**

Keywords-workload balance; TSP; KTSP; 3-opt; genetic; medium-sized

I. INTRODUCTION

KTSP(K-person Traveling Salesman Problem) is an extension of TSP(Travelling Salesman Problem) problem, which is used to solve some path planning problems in real life, and has a wide range of applications. For example, software companies should send a number of technicians to maintain the software systems purchased by customers all over the country, fast-food enterprises have delivery problems of takeout with several distributors, advertising companies' leaflet delivery, etc. These real application problems in life can be attributed to KTSP problems. TSP and KTSP are NP(Non-deterministic Polynomial) complete problems[1,2]. The optimal solution of the problem can be solved accurately by cutting plane algorithm, but the algorithm has the disadvantage of high time complexity, so it becomes very difficult to solve medium and large-scale problems[3,4,5]. Solving TSP problem with heuristic method is a kind of method widely studied at present, mainly including greedy Algorithm[6], genetic algorithm [7], simulated annealing algorithm[8],k-opt algorithm, LK algorithm, LKH algorithm[9-11],ant colony algorithm [12], particle swarm optimization algorithm[13], etc. These methods are mainly based on the idea of intelligent optimization, so they have fast search ability and can obtain high-quality solutions[14-15]. They are worthy of reference for KTSP planning with large scale. Generally, the KTSP problem basically does not consider the average workload between each salesman, so it is basically similar to the TSP planning method, and it is easy to find the optimal solution, but it is difficult to be applied to practice. The research content of this paper fully considers the workload balance of each salesman. Firstly, KTSP is transformed into TSP with workload balance problem, then hunting method and 3-opt(3-optimization) algorithm are used to search the local optimal solution, genetic mutation method is used to search the global optimal solution, and the workload balance algorithm is used to correct the results, which has high ability to solve the optimal solution, It is effective for medium-scale equilibrium KTSP problem.

II. BALANCED KTSP PROBLEM

For the general KTSP problem, given k salesmen and n cities, k salesmen start from city 0 and visit the other n-1 cities. Each city needs and only needs one salesman to visit once, and finally all return to the starting city 0. The goal is to plan the line with the shortest sum of all marketing paths. When studying practical problems, this paper requires that the workload of each salesman should be balanced, that is, the number of cities visited by each salesman should be the same, so as to eliminate the problem that the workload of individual salesmen is too small and the workload of individual salesmen is too large in the planning results, which forms a KTSP problem with workload balance. KTSP problem can be described in the form of graph theory: let V=(v_1,v_2, ...,v_n)is a set of n points on the plane, G= (V, E) is a complete graph on vertex set V and edge set E. Let H be the set of k sub paths in graph G = (V, E) starting from city v_1 and returning to v_1 after cruising multiple cities, which is recorded as H = (H_1, H_2,..., H_k). Count (V) is used to represent the number of elements in vertex set V, and the constraints are (1), (2) and (3).

$$v_i \neq v_j, (v_i \in \mathrm{H}_a, v_j \in \mathrm{H}_b, i > 1, j > 1) \quad (1)$$

$$\mathrm{V} = \cup\, \mathrm{V}_{Hi}\ ,(\mathrm{Hi} \in \mathrm{H}) \quad (2)$$

$$\mathrm{Count}(\mathrm{V_{Hi}}) = \mathrm{Count}(\mathrm{V_{Hj}}), (\mathrm{Hi} \in H, \mathrm{Hj} \in H) \quad (3)$$

In other words, except for the starting point, there are no same vertices between the sub paths. Any vertex will be passed by the cruise path once and only once, and the number of vertices contained in any two sub paths is the same. The KTSP problem is to find the shortest k travel path H.

E_{ij} represents the length of the edge from vertex i to vertex j; If the edge from vertex i to vertex j is included in the cruise path, it is recorded as δ_{ij} = 1, otherwise δ_{ij} =0.

The objective function for finding the minimum k-cycle path length can be expressed as (4) and (5).

$$f(V) = min \sum_{i \in V, j \in V} \delta_{ij} E_{ij} \quad (4)$$

$$\text{among,}\ \ \delta_{ij} = \begin{cases} 1 & Salesman\ passes\ i,\ \mathrm{j} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

III. ALGORITHM DESCRIPTION:

Since only the local optimal solution can be obtained by optimizing the sub path alone, we can form all the sub paths

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00044

into a whole path and then optimize the whole path, so as to obtain the global optimal solution. Considering that each salesman starts from the starting city and finally returns to the starting city, and the sub paths are connected to form a closed loop that passes through the starting city many times, we can virtual k-1 vertices that completely coincide with the starting city, and set the distance between the K coincident starting cities as infinity. In this way, we can solve the KTSP path problem through global optimization. Let the set of all k starting cities be recorded as C, then the distance between cities can be expressed as (6):

$$d_{ij} = \begin{cases} \infty, & i \in C, j \in C \\ \sqrt{(x_i - x_j)^2} + \sqrt{(y_i - y_j)^2}, & (other) \end{cases} \quad (6)$$

Through the nearest neighbor algorithm, an initial feasible solution of workload balance is found and saved as the current optimal solution.

A. *global planning using genetic mutation algorithm*

Input: optimalSolution:the current optimal solution
Output: optimalSolution:the new optimal solution
1) while true do
2) testSolution ← optimalSolution;
3) testSolution ← geneticAlgorithm(testSolution)[7]
4) testSolution ← huntingAlgorithm(testSolution)
5) testSolution ← 3-optAlgorithm(testSolution)[10]
6) if testSolution is worse than optimalSolution then
7) exit
8) end
9) testSolution ← workBalanceAlgorithm(testSolution)
10) if testSolution is better than optimalSolution then
11) optimalSolution ← testSolution
12) end
13) end

B. *local optimization algorithm*

In the loop sequence, if a point in the line is captured by the edge close to it, the position of the point changes, which shortens the length of the whole line. We call this algorithm hunting algorithm. As shown in Fig. 1, point E in line AB is too close to line CD, and point E is captured by line CD, thus shortening the length of the whole line.

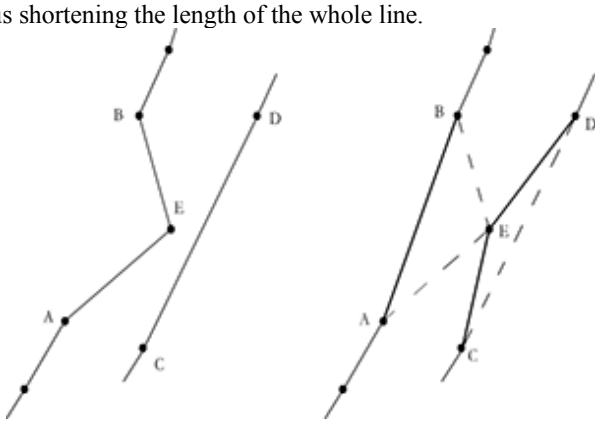


Figure 1. hunting method

In the process of local optimization, the 3-opt algorithm[10] is adopted. The algorithm uses three new edges to replace the three edges on the current loop to obtain a new loop. If the new loop solution after exchange is better, the result will be adopted, otherwise the attempted loop solution will be abandoned. As shown in Fig. 2, it shows that CD edge, EF edge and AB edge are replaced by the other three edges.

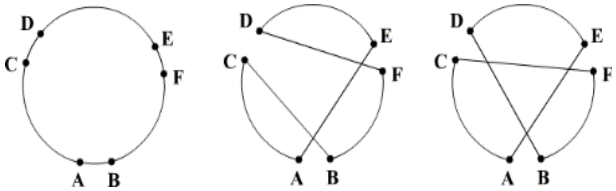


Figure 2. 3-opt algorithm

C. *workload balancing algorithm*

According to the results of the previous planning, the workload of each salesman may be different. In order to average the workload of each salesman, the following workload balancing algorithm is used:

Input: *cruisePath* with unbalanced workload
Output: *cruisePath* with balanced workload
1) for each *node* in *cruisePath* do
2) if *node* = *centralNode* then
3) connect the front and rear nodes of *centralNode*
4) delete *centralNode*;
5) end
6) end
7) start a new *subPath*;
8) for each node in *cruisePath* do
10) add the *node* to the *subPath*
11) *nodesNumber* ← the nodes number of the *subPath*
12) if *nodesNumber* = *workload* then
13) add a *centralNode* after the *node*
14) start a new *subPath*
15) end
16) end

After the workload balancing algorithm, if the total length of the planned path is less than the current optimal solution, a better solution is obtained, and the better solution is saved as a new optimal solution.

Taking VS2010 as the development tool, the author programmed and implemented the algorithm described in this paper. From the test results, the algorithm is effective and the effect is ideal. This algorithm can find the global optimal solution for medium-sized balanced KTSP path planning. The algorithm considers the workload balance of each salesman globally, so it can better balance various factors in path planning. At the same time, the heuristic algorithm 3-opt algorithm with the most outstanding performance is used to obtain the local optimal solution, search the global optimal solution through genetic mutation, and distribute the same workload through workload balance.

IV. ALGORITHM OPERATION EFFECT

When Huang Xiyue studied the non load unbalanced KTSP problem [1], he adopted the standard data set tsplbi95. (http://comopt.ifi.uni-heidelberg.de/software/TSPLIB95/). He studied the example eil51 of TSPLIB95, but the number of urban nodes is relatively small, only 51. The workload balanced KTSP studied in this paper is more difficult than the KTSP problem, but in order to further study the operation effect of the algorithm and deal with more cities, this paper selects another larger example eil101 in the

TSPLIB95 dataset Test and research. For the solution of KTSP problem, in order to facilitate comparison with TSP algorithm, we set the simplified condition that the demand of each vertex is 1 and the workload of each salesman is the same. Eil101 in the data set is a complete graph of 101 cities. This paper sets the starting point at the central 101 node and tests the lines when the number of salesman is k = 2, 4 and 5 Road planning. Compared with reference [1] and reference [2], the balanced workload KTSP algorithm studied in this paper has the advantages of larger data processing scale and better effect.

For eil101 data set, the planning algorithm is run separately for 10 times, 30000 iterations are executed each time, and the planning results are recorded. The worst solution, optimal solution and average value are shown in Table I. It can be seen that in a certain range, with the increase of the number k of salesmen, the value of the optimal solution also increases.

TABLE I. RESULTS OF EIL101 OF 101 NODES RUNNING 10 TIMES IN EACH CASE WHEN K TAKES DIFFERENT VALUES

K	Workload of sub Tour	Optimal solution	Worst solution	average value
2	50	656.8	672.5	660.4
4	25	714.7	796.7	751.8
5	20	735.6	847.1	790.2

During the planning process, record the change process of the optimal solution, as shown in Fig.3. The evolution curve shows that the optimal value decreases rapidly at the initial stage of optimization. With the increase of optimization times, the decline of the optimal value slows down and finally tends to the level. It can also be noted that the descent curve is the steepest when k = 2 and the most gentle when k = 5. The results show that in a certain range, with the increase of the number k of salesmen, the planning becomes more and more difficult.

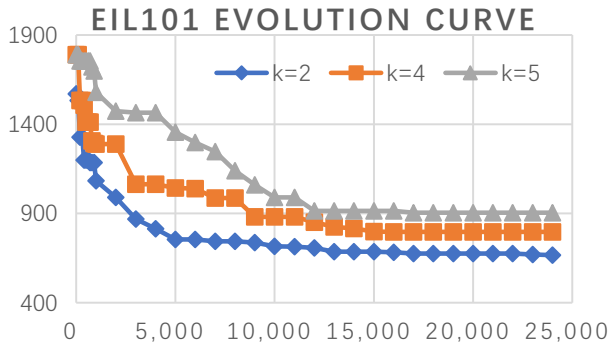


Figure 3. Evolution curve of eil101 at k = 2, 4 and 5

Using this algorithm to solve the eil101 problem, when k = 2, the optimal route is shown in Fig. 4, and the length value of the optimal solution is 656.9. When k = 4, the optimal route is shown in Fig. 5, and the optimal solution is 714.7. When k = 5, the solution result is shown in Fig. 6, and the optimal solution is 735.6. The details of the specific city access sequence when k takes different values are shown in Table II.

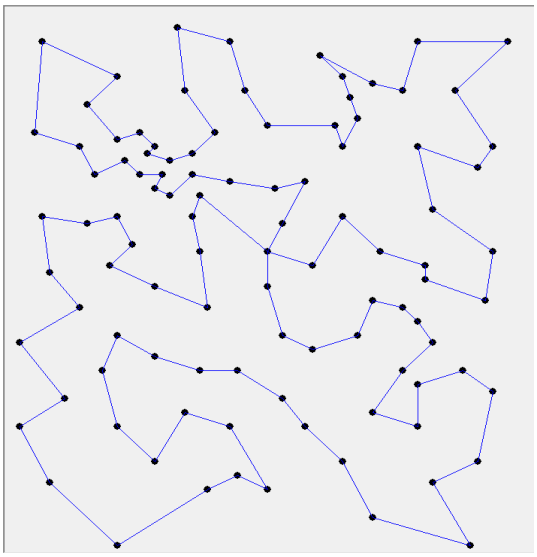


Figure 4. TSPLIB95 example eil101, when the number of operators k = 2, the optimal solution path

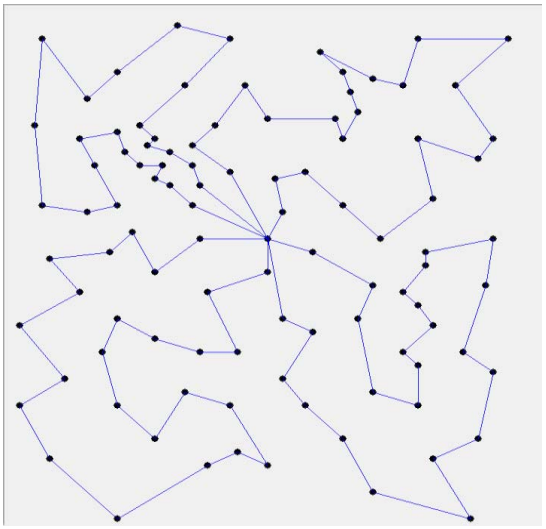


Figure 5. TSPLIB95 example eil101, optimal solution path when the number of operators k = 4

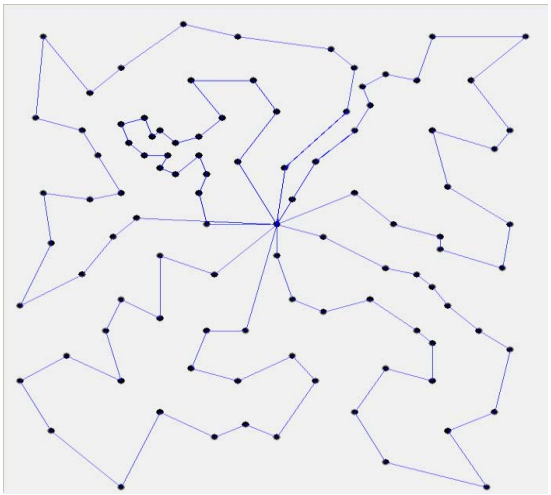


Figure 6. TSPLIB95 example eil101, when the number of operators k = 5, the optimal solution path

With the increase of the number of salespeople and K value, the demand for computing power increases, and the length of the minimum cruise path solution tends to increase. In particular, when k = 1, it is an ordinary TSP problem.

TABLE II. COMPARISON OF OPTIMAL PATHS OF EXAMPLE EIL101

K	optimal value	Access path order
2	656.8	101,94,6,89,52,18,83,60,5,84,17,45,8,46,47,36,49,64,63,90,32,10,62,11,19,48,82,7,88,31,70,30,20,66,65,71,35,34,78,81,9,51,33,79,3,77,76,50,1,69,27, 101,53,40,58,13,95,96,99,59,93,85,61,16,86,38,14,44,91,100,37,98,92,97,87,42,43,15,57,2,73,21,72,74,22,41,75,56,23,67,39,25,55,4,54,24,29,68,80,12,26,28,101
4	714.7	101,69 ,1 ,70 ,30 ,20 ,66 ,65 ,71 ,35 ,34 ,78 ,29 ,24 ,80 ,68 ,77 ,3 ,79 ,33 ,81 ,9 ,51 ,50 ,76 ,28 , 101 ,53 ,58 ,40 ,26 ,12 ,54 ,4 ,55 ,25 ,39 ,67 ,23 ,56 ,75 ,41 ,22 ,74 ,72 ,21 ,73 ,2 ,57 ,87 ,97 ,13 , 101 ,94 ,95 ,92 ,98 ,37 ,100 ,42 ,15 ,43 ,14 ,44 ,38 ,86 ,17 ,84 ,5 ,61 ,16 ,91 ,85 ,93 ,59 ,99 ,96 ,6 , 101 ,89 ,18 ,60 ,83 ,45 ,8 ,46 ,47 ,36 ,49 ,64 ,63 ,90 ,32 ,10 ,62 ,11 ,19 ,48 ,82 ,7 ,88 ,31 ,52 ,27 ,101
5	735.6	101,52 ,18 ,7 ,82 ,48 ,19 ,47 ,36 ,49 ,64 ,11 ,63 ,90 ,32 ,30 ,70 ,10 ,62 ,88 ,31 , 101 ,28 ,76 ,77 ,3 ,79 ,78 ,34 ,35 ,71 ,65 ,66 ,20 ,51 ,9 ,81 ,33 ,50 ,1 ,69 ,27 , 101 ,58 ,73 ,22 ,41 ,15 ,43 ,14 ,44 ,38 ,86 ,16 ,61,5 ,84 ,17 ,45 ,46 ,8 ,83 ,60 , 101 ,89 ,6 ,94 ,95 ,96 ,99 ,59 ,93 ,85 ,91 ,100 ,98 ,37 ,92 ,97 ,87 ,42 ,57 ,2 ,13 , 101 ,53 ,40 ,21 ,72 ,74 ,75 ,56 ,23 ,67 ,39 ,25 ,55 ,4 ,54 ,24 ,29 ,68 ,80 ,12 ,26 ,101

V. ANALYSIS OF EXPERIMENTAL RESULTS

A large number of experiments show that the optimal solution of the model studied in this paper has some characteristics, which is basically distributed in a flower shape around the distribution center. If the nodes are sparse, the patrol paths of each salesman will not intersect. If the nodes are dense, the petals may intersect near the distribution center. Each petal may contain another petal under certain circumstances. In the same petal, there is no path crossing. The study of such optimal solution characteristics has guiding significance for the further development of high-performance algorithms.

After considering the workload balance, the difficulty of solving KTSP is greatly increased. The convergence speed of the method given in this paper is still slow for the large-scale balanced KTSP problem. Through the analysis, we can know that the problem is that when planning to obtain the better value, it does not necessarily meet the conditions of workload balance. This two-step planning method has limitations. The next research can combine the solution of the better value with workload balance to improve the efficiency of the algorithm.

VI. CONCLUSION

In this paper, genetic algorithm, 3-opt algorithm, hunting algorithm and balanced workload algorithm are used to solve the workload balanced KTSP problem. Firstly, KTSP is transformed into TSP problem through virtual central node, and the local optimal solution is obtained by using 3-opt algorithm and hunting algorithm. Then, the result is corrected by workload balancing algorithm, and finally the optimal solution is searched globally by genetic algorithm. The practical results show that the algorithm is effective in solving the medium and large-scale equilibrium KTSP problem.

REFERENCES

[1] HUANG Xi-yue 1 , HU Xiao-bing, Application of Ant Colony Algorithm in K-person Traveling Salesman Problem, Journal of computer simulation, 2004(12),pp.162-164

[2] WENG Miao-feng LI Wen-yu, Applying Chaotic Neural Network Based on Annealing Strategy to KTSP Problem, Journal of East China Shipbuilding Institute(Natural Science Edition) 2003(08),pp.42-46

[3] Manfred Padberg,Giovanni Rinaldi,A Branch-and-Cut Algorithm for the Resolution of Large-Scale Symmetric Traveling Salesman Problems,SIAM REV, Mar 1991

[4] Kahng, A.B.. (1989). Traveling salesman heuristics and embedding dimension in the Hopfield model. 513 - 520 vol.1. 10.1109/IJCNN.1989.118627.

[5] Naddef D. (2007) Polyhedral Theory and Branch-and-Cut Algorithms for the Symmetric TSP. In: Gutin G., Punnen A.P. (eds) The Traveling Salesman Problem and Its Variations. Combinatorial Optimization, vol 12. Springer, Boston, MA.

[6] G.Gutin and A.Yeo.The Greedy Algorithm for the Symmetric TSP.Algorithmic Oper. Res., Vol.2, 2007, pp.33--36. // Greedy

[7] P. Merz and B. Freisleben, "Genetic local search for the TSP: new results," Proceedings of 1997 IEEE International Conference on Evolutionary Computation (ICEC '97), 1997, pp. 159-164, doi: 10.1109/ICEC.1997.592288.

[8] Xiutang Geng, Zhihua Chen, Wei Yang, Deqian Shi, Kai Zhao,Solving the traveling salesman problem based on an adaptive simulated annealing algorithm with greedy search,Applied Soft Computing,Volume 11, Issue 4,2011,Pages 3680-3689

[9] G. A. Croes, A Method for Solving Traveling-Salesman Problems,Operations Research, INFORMS, vol. 6(6), 1958(12),pp.791-812.

[10] K. Helsgaun,General k-opt submoves for the Lin-Kernighan TSP heuristic.Mathematical Programming Computation,1(2-3):119-163 (2009).

[11] Tinós R., Helsgaun K., Whitley D. (2018) Efficient Recombination in the Lin-Kernighan-Helsgaun Traveling Salesman Heuristic. In: Auger A., Fonseca C., Lourenço N., Machado P., Paquete L., Whitley D. (eds) Parallel Problem Solving from Nature – PPSN XV. PPSN 2018. Lecture Notes in Computer Science, vol 11101. Springer, Cham. https://doi.org/10.1007/978-3-319-99253-2_8

[12] S Li, S Cai, L Li, R Sun, G Yuan,CAAS: a novel collective action-based ant system algorithm for solving TSP problem, Soft Computing,2020

[13] Wang Qingxi;Guo Xiaobo,Particle swarm optimization algorithm based on Levy flight, Application Research of Computers, 2016,33(9):2588-2591

[14] V Gabrel , AR Mahjoub , R Taktak , E Uchoa,The Multiple Steiner TSP with order constraints: complexity and optimization algorithms, December 2020,Soft Computing 24(3)

[15] UshaMohan,SivaramakrishnanRamani,SounakaMishrac,A 4-approximation algorithm for the TSP-Path satisfying a biased triangle inequality, Discrete Applied Mathematics Volume 271,1 December 2019, Pages 108-118

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

An Intelligent Identification and Detection Method for Forest Pests Based on YOLOv5s Algorithm

Zhihong Wang[1],Zhanhao Shi[1],Honglei Wang[2],Fengju Bu[1],Tian Liang[1],Xingbo Zhang[1]
[1]Shandong Agriculture and Engineering University
[2]Jinan Landscape and Forestry Science Research Institute
Jinan,China

***Abstract*—Accurate identification of pests is the basis of forestry pests prevention and management. This paper uses the YOLOv5s algorithm to train the pest intelligent identification model by using the cosine annealing learning rate to adjust the learning rate during the training process.The model can identify 15 common forestry pests.When their different physiological forms such as eggs, larvae,pupae and adults are concluded,the model can identify a total of 39 categories accurately.**

***Key words*—Deep learning;YOLOv5s;Forestry pest detection**

I.INTRODUCTION

Forestry pest control is an extremely important content in forestry construction. By the end of 2020, my country's forest coverage rate has reached 23.04%.Our country forest area is very small, but it is one of the countries where forestry pests and diseases occur more severely in the world [1]. There are more than 200 kinds of forestry pests that often cause harm,among which more than a dozen kinds are particularly serious, such as pine wood nematode, pine caterpillar, American white moth, pine ink longicorn beetle, yellow spotted longicorn beetle and so on. Large-scale outbreaks of pests be avoided as early as possible only by accurately and timely monitoring and killing forest pests.To achieve this goal, we must first realize the accurate identification and classification of forestry pests.

The traditional pest identification method is generally based on the shape characteristics of pests by experts and technicians, comparing with the accurately recorded type specimens to give the types of pests, whose efficiency and accuracy are limited. In recent years, computer vision and deep learning models have developed rapidly, and researchers have also applied these technologies to the field of pest detection and control. Li Wenyong et al.[2]proposed an algorithm to identify automatic trap target pests through representing and learning the pose features of orchard pests. CHENG X et al. [3] built a new convolutional neural network model for pest identification by using the residual module in RESNET for reference and deepening the network on the basis of Alex net.Zhou Aiming [4] realized the identification and counting of rice pests based on deep learning technology, and the recognition accuracy was about 90%. LINTL et al.[5] realized pests detection in the growth of sweet peppers based on the Faster R-CNN method. Li Hengxia et al.[6]proposed a rape pest detection better adapt to different data sets based on deep convolutional neural network.

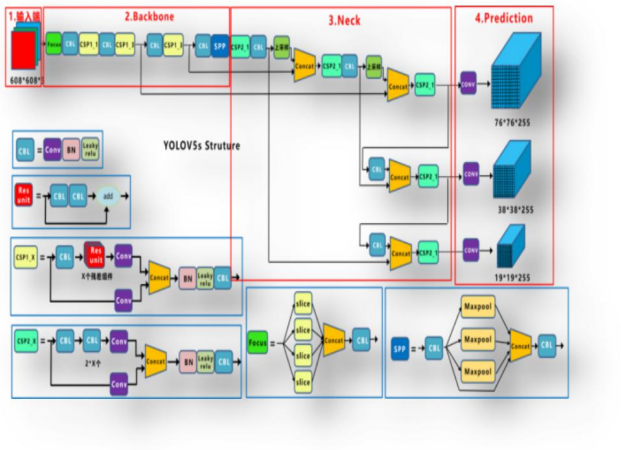
In this paper, the algorithm YOLOv5s is applied to the field of forestry pest detection.And it is the latest first-order target detection algorithm with fast detection speed.In the training process, the cosine annealing learning rate method is used for model training, and the Focus layer of YOLOv5s is improved. The trained model classifies 39 kinds of pests, and the test accuracy on the training set is as high as 98.6%, The correctness of the test set is 94.5%, and the speed is improved.

II.YOLOV5 MODEL INTRODUCTION

The YOLO algorithm is a new target detection framework developed by Redmon et al. [7] and it can convert the target detection problem into a regression problem[8]. The YOLOv5 algorithm is the latest version.Currently,there are four models such as v5x, v5l, v5m, and v5s. The smallest v5s model is only 27MB, which is only 11% of the YOLOv4 model. The speed reaches 140 frames per second.It is currently the fastest algorithm in target detection[9-12].

Compared with the original YOLOv3 and YOLOv4 framework, the YOLOv5 model has been improved in data enhancement, selection of loss function, and selection of activation function. On the activation function, YOLOv5 algorithm selects Leaky ReLU and Sigmoid as the activation function replacing the Swish function in YOLOv4. In terms of loss function, YOLOv4 algorithm selects CIOU as the coordinate value loss function, while YOLOv5 algorithm selects GIOU as the coordinate value loss function. YOLOv5 adaptively learns the initial anchor box size before each detection, which can better adapt to different data sets.The network model structure is shown in Figure1.

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00045



FigureJ YOLOV5 model network structure

III. EXPERIMENTS AND RESULTS

A. Data set production

Data were collected through web crawlers and open source data sets, etc.We have collected 15 common pests and their different physiological forms.And finally we get a total of 39 categories pests. The actual training and test data sets select 1357 pictures.Data sets are shown in Figure 2.

Figure2 Training samples

B. Data processing

The data set expansion is realized by using OpenCV Python tools based on image radiation transformation, noise distribution, horizontal mirror and other operations.An example of data enhancement is shown in Figure3

Figure3 Data enhancement processing

After the data is expanded, the 15 pest species and the number of samples for each type are expanded as shown in Figure4. The original 1357 pictures have been expanded to 33000, and a total of 27000 label objects are included.Training sets and test sets are divided by 8:2.

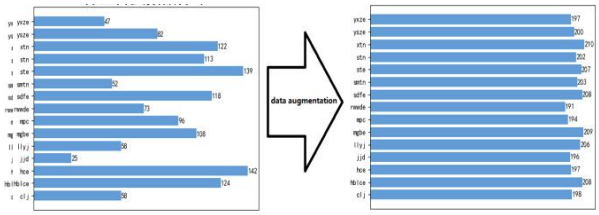


Figure4 Data distribution after enhancement

C.Picture Annotation

The annotation interface with LabelImg software is shown in Figure5 and Figure6.

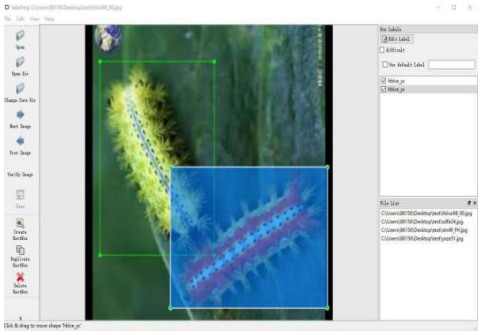


Figure 5 Annotated brown-edge green spiny moth-larva

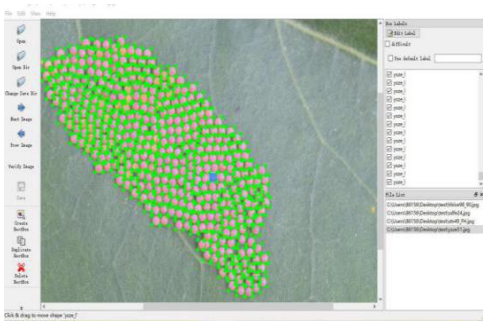


Figure6 Annotated Yang Fanzhou Moth-Egg

D. Model Training

- Experimental Environment

The experimental environment selects the pytorch architecture.CPU is i7-11800H, and the graphics card is RTX3070.

- Model Evaluation

This paper uses the cosine annealing learning rate to adjust the learning rate. The cosine annealing learning rate is different from the traditional learning rate. With the increase of epoch, the learning rate first drops sharply, then rises sharply, and then this process is repeated continuously.The experimental results show that the yolov5s algorithm using the cosine annealing fully meets the needs of fast and accurate use. The training loss after the algorithm optimization has a shock phenomenon at the beginning, but then Loss tends to steadily decrease. As shown in Figure7.

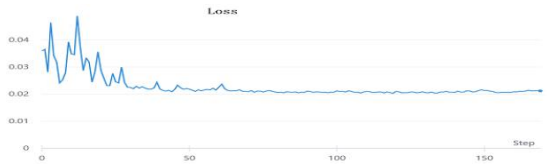


Figure7 Loss1

The convergence of the training model after the first 170 epochs tends to be flat. In response to this situation, a new training set and a test set are re-adjusted and divided, and the learning rate is reduced by 10 times, and the second round of training is carried out. After the second round of training, Loss tends to 0. As shown in Figure8 below.The test accuracy of pest identification rate on the training set is as high as 98.6%, and the test set is about 94.5%. So far, the model training is completed.

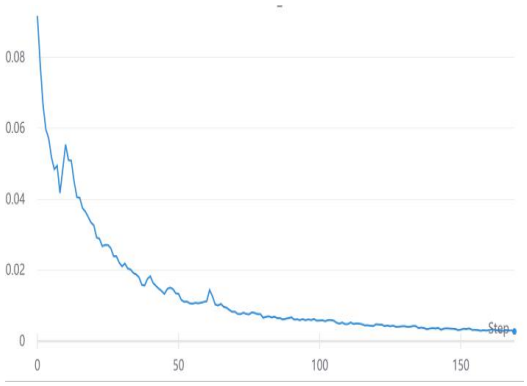


Figure8 Loss2

IV.CONCLUSIONS

In summary, it is proposed to use YOLOV5s network model to identify and classify forest pests. This method has been experimentally tested and has a high accuracy rate. However, because the training data and the user test data are taken in different ways, it has a certain impact on the accuracy. The transfer learning can be made on the existing basis to make the algorithm more accurate. The recognition speed and accuracy can be dynamically adjusted according to the computing performance of the terminal device.

REFERENCES

[1] WANG Zhenghong. Research on recognition and classification method of forestry pests based on fractal theory[D].Ha'er bin: Northeast Forestry University,2010.

[2] Li Wenyong, Li Ming, Chen Meixiang, et al. Feature extraction and classification method of crop multi-pose pests based on machine vision [J]. Journal of Agricultural Engineering, 2014, 30(14):154-162.

[3] CHENG X,ZHANG Y,CHEN Y,et al. Pest identification via deep residual learning in complex background[J].Computers and Electronics in Agriculture,2017,141: 351-356.

[4] Zhou Aiming.Automatic recognition and counting technique for agricultural lamp-trap pests based on deep learning[D].Hangzhou: Zhejiang Sci-Tech University,2018.

[5] LINTL, CHANG H Y, CHEN K H. Pest and disease identification in the growth of sweet peppers using faster R-CNN[C].2019 IEEE International Conference on Consumer Electronics,2019.

[6] LI Hengxia, LONG Chenfeng, Zeng Meng,et al. A detecting method for the rape pests based on deep convolutional neural network [J].Hunan Agricultural University Natural Sciences,2019, (5):560-564.

[7] Redmon J,Divvala S,Girshick R,et al. You Only Look Once:Unified,Real-Time Object Detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE Press, 2016: 779-788.

[8] GAO Chunyan, ZHAO Wenhui, ZHANG Minglu, et al. A detectiong method for the rape pests based on deep convolutional neural network[J]. Journal of Tianjin University(Science and Technology), 2020, 53(4): 358-365.

[9] Hou Ruihuan, Yang Xiwang, Wang Zhichao, Gao Jiaxin. A real-time detection method of forest pests based on YOLOv4-TIA. Computer Engineering. https://doi.org/10.19678/j.issn.1000-3428.0060563

[10] Tan Shilei, Bie Xiongbo,Lu Gonglin,Tan Xiaohu. Realtime detection of masks worn by people based on the YOLOv5 network model[J].LASER JOURNAL.2021,42(2):147-151.

[11] Bochkovskiy A,Wang C Y,Liao H Y M.YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. 2020,57 (5) :9-12.

[12] Wu D,Lv S,Jiang M,et al.Using channel pruning-based YOLOv4 deep learning algorithm for the realtime and accurate detection of apple flowers in natural environments[J].Computers and Electronics in Agriculture,2020,178 (5): 174-178

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Improve the Performance of LSM-Tree based on Key-Value via Multithreading

Yuan Gao[1,2,3,4] , Ping Xie[1,2,3,4,*], Wendi Hua[1,2,3,4], Meng Lv[1,2,3,4], Jiating Lu[5]
1. College of Computer of Qinghai Normal University, Xining 810016, P. R. China
2. The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Xining 810016, P. R. China
3. Key Laboratory of Internet of Things of Qinghai Province, Xining 810016, P. R. China
4. Academy of Plateau Science and Sustainability, Xining 810016, P. R. China
5. School of Physical and Electrical Engineering，Qinghai Normal University，Xining 810016, P. R. China
*Corresponding author: xieping@qhnu.edu.cn

***Abstract*—In the era of big data, key-value storage systems based on Log-Structure Merge tree (LSM-tree) are widely used in numerous industries. LSM-Tree is divided into two parts, one part is in the memory and the other part is on the hard disk. When writing data, first write to the memory, when the Immutable Memtable in the memory is full of data, then write to the disk to output a Sorted String Table (SSTable) file. However, the use of writing data from memory to disk is single-threaded. When the Immutable Memtable writes data to the disk, it will block other threads. This article uses multiple Immutable Memtables and multi-threaded writing to solve the single-thread blocking problem. By running the benchmarks of LevelDB to analyze the experiment, the experimental results indicate a higher improvement.**

Keywords—Key-Value store, Multi-Threading, Write performance, Read performance , LSM-Tree, Immutable Memtable

I. INTRODUCTION

With the advent of big data, data is the most important asset of data centers and even enterprises. In the data society, data has the dual role of basic strategic resources and key production factors. As the core part and underlying base of the information system, the construction and use of storage systems are directly related to the storage, use, and mining of data, which is the core asset of the enterprise. Currently, the challenges faced by storage systems are as follows. First, the amount of data has exploded. With the continuous development of the mobile Internet, the scale of enterprise data has shown explosive growth. Research shows that in 2018, China's new data will be 7.6ZB; in 2025, China's new data will reach 48.6ZB, with an average annual growth rate of 30%. Second, the business is responsible for presenting dynamic changes. A load of modern service platforms is non-linear and dynamically changing, especially for Internet-based services, and sudden changes in service load may occur at any time. Taking Double Eleven in 2020 as an example, the real-time transaction volume of online shopping malls exceeded 372.3 billion yuan. The peak order creation is 583,000 transactions per second, which is 1457 times that of the first Double Eleven event in 2009.

Currently, key-value storage has become an important part of various data-intensive storage applications. Compared with traditional storage structures, non-traditional storage has many advantages, including scalability, ease of use, and higher throughput. Non-relational databases are broadly divided into four categories based on data size: key-value, wide-column, document, and graph. Typical applications of key-value storage include graph databases, task queues, stream processing engines, NoSQL storage, and distributed databases. Increasing demand is higher reading and writing performance. According to the current situation, workloads increasingly tend to be key-value storage. Therefore, it is necessary to optimize key-value storage.

The current mainstream key-value storage structure is LSM-Tree[1], which is a hard disk-based data structure. Compared with B-Tree, it can significantly reduce the overhead of the hard disk and provide a longer time for high-speed insertion. LSM-Tree divides the storage structure into two parts, one part in the memory and the other part in the hard disk, which can make good use of the high-speed performance of the memory and the large capacity of the disk. Using such a model can trade high-speed writing performance at the expense of small read performance. The data in memory is stored in the Memtable and immutable Memtable in order. In the disk, the structure for storing data is SSTable. The memory and data in SSTable are ordered and unique. The one exception is data in disk tier 0. Because all memory data needs to be received at a high speed, the data at level 0 is repeated. The data storage process of LSM-Tree is generally written to the Memtable first, then to the immutable Memtable when it is full, and then flush to level 0 of the disk when it is full. After that, after level 0 is full, it is compressed and merged to level 1, and so on.

Many articles improve performance by optimizing the data structure of LSM-Tree. Optimize the SkipList memory data structure and optimize the disk storage structure SSTable to improve writing and reading performance. This paper analyzes the LSM-Tree writing process and analyzes each module one by one to find the optimization points. Through analysis, it is found that when the immutable Memtable is flushing data to the disk, LSM-Tree uses a single thread for flushing. When the immutable Memtable is inputting data, it will block the Memtable from writing data to the Immutable Memtable, and by reading the paper SIKL[7], it is found that LSM-Tree will trigger the long tail delay in three places, of which the immutable Memtable flush to the disk is a very important trigger point. And get inspiration from the article Multiple Immutable Memtables, which increases efficiency by increasing the number of immutable Memtables and writing multiple immutable Memtables to disk at the same time through multithreading.

II. BACKGROUD

A. Concept of LSM-Tree

Fig 1 shows the basic structure of the original LSM-Tree. LSM-Tree is divided into two parts, one part in the memory and the other part in the hard disk. The part on the hard disk

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00046

is divided into multiple parts, each part is called a level, from top to bottom is level 0 to level n. When writing data, first write it to C_0 of the memory. When C_0 is full, C_0 will flush the data to the disk, and C_0 will flush the data to c1. When the data of c1 is full, it will be compressed and merged with C_2, And so on, until the C_n layer. As the number of layers increases, the capacity of each layer will increase exponentially. Because the data is continuously compressed downwards, the new data is in the lower layer and the old data is in the higher layer.

Fig 2 shows the basic structure of LevelDB[2]. LevelDB optimizes the structure of LSM-Tree. The memory c0 is divided into two parts, one part is Memtable, and the other part is immutable Memtable. Created a new data structure named SSTable, each layer contains many SSTables. For some auxiliary functions, log files, manifest files, and current files are introduced. The following briefly introduces the function of each part.

Memtable: Memory data structure, SkipList implementation, new data will be written here first.

Log file: Before writing to Memtable, the log file will be written first, and the log will be written sequentially by append. Another function of logs is that they can be used to restore data after the machine is down.

Immutable Memtable: After reaching the upper limit of the capacity set by the Memtable file, the Memtable will be converted to an immutable Memtable. This is to prepare for the conversion to an SSTable file. The immutable Memtable is not allowed to be modified. After the immutable Memtable is written, a new Memtable will be generated.

SSTable file: Disk data storage file. From level 0 to level n of the disk, each layer contains multiple SSTable files. As the number of layers increases, the total amount of SSTable contained in each layer increases exponentially, and the data in the SSTable is ordered . Among them, the SSTable file in level 0 is generated by directly dumping the immutable Memtable of the memory, and the other level SSTable files are generated by merging the files of the upper layer and the files of this layer. The SSTable files are written and generated sequentially during the merging process. Later, it can only be deleted in subsequent merges without any modification operations.

Manifest file: The manifest file records the distribution of SSTable files at different levels, the maximum and minimum keys of a single SSTable file, and other meta-information required by LevelDB.

Current file: When LevelDB starts, the first task is to find the current manifest file, and there will be multiple manifest files, the current file will record the file name of the current manifest, making it easy to find.

After introducing each component of LevelDB, let's briefly introduce the reading process and writing process. Writing process: The written operation of LevelDB includes setting key-value and deleting key. It should be pointed out that these two cases are the same in the processing of LevelDB. The deleted operation is actually to insert a piece of data marked as deleted into LevelDB. The external write interfaces provided by LevelDB include put, delete and write. Among them, writing needs a WriteBatch as a parameter, and put and delete first encapsulate the current operation into a WriteBatch object and call the write interface. A WriteBatch is a collection of a batch of write operations, and its significance is to improve writing efficiency and provide atomicity of all writes in the batch. In the write function, a writer is first encapsulated with the current WriteBatch, which represents a complete writing request. The LevelDB lock ensures that only one writer can work at the same time. Other writers suspend and wait until the previous writer finishes executing and wakes up.

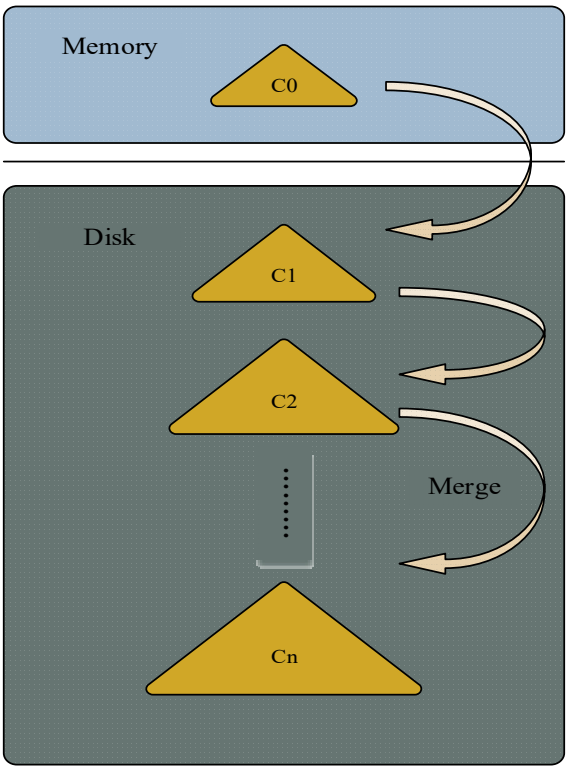


Figure 1: The original structure of LSM-Tree

The SSTable file on the disk is also very important. The keys in the SSTable file are ordered, and the key value in the file is the smallest. This is the case for each level of SSTable. However, the SSTable at level 0 is different from the SSTable at other levels and the key ranges of the two SSTable files at level 0 overlap. A certain file in SSTable belongs to a specific level, and the keys are in order, so it is very important to record the smallest key and the largest key. The manifest file is used to store this information. It records the management information of each file in SSTable, Such as which level it belongs to, what is the file name, and what is the minimum and maximum key.

Reading process: First, generate the key used for the internal query, which is generated by splicing the sequence with the UserKey requested by the user. The sequence can be provided by the user or use the current latest sequence, and LevelDB can ensure that only the writes before this sequence are queried. Using the generated key, try to read from the Memtable, immutable Memtable, and SSTable files in turn until it is found. Finding from the SSTable file needs to try to read in each layer in turn, because the key

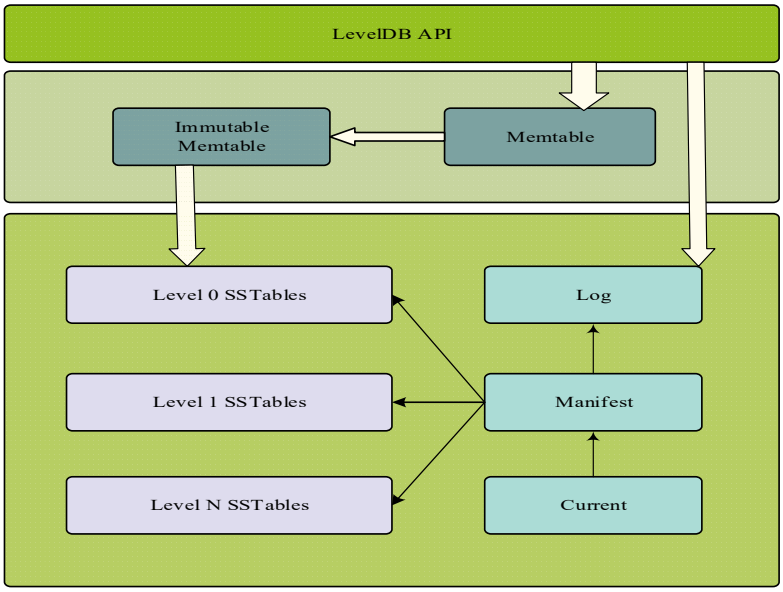


Figure 2:The structure of LevelDB

interval of each file recorded in the manifest can quickly know whether a key is in the file. Since level0 files are directly dumped from immutable Memtable, they will inevitably overlap with each other, so each file needs to be searched once. For other levels, because the merging process ensures that they do not overlap and are in order, the binary search method provides better quality efficiency.

B. Related Work

By studying the overall structure and read-write process of LSM-Tree, the current article is working on the following six aspects, such as: write amplification, merge operation, hardware, special workloads, Auto Turning, and Secondary Indexing. The write amplification problem is the root problem of LSM-Tree. There are many articles on optimizing this problem, such as: LWC-tree[8], PebblesDB[3], dCompaction[4], SifrDB[5], TRIAD[6], etc. Changing the compression method of each layer of SSTable on the disk can also improve the efficiency of LSM-Tree. There are the following research results: bLSM[10]. At present, the emergence of new storage media also provides a good opportunity for the application of LSM-Tree. The following articles are optimized under the new storage media: WiscKey[11], etc. In the new application environment, its special load has also changed. Like some extreme loads, one hundred thousand or one million visits will be generated within a few milliseconds. In response to these extreme loads, the following results have been produced: SlimDB[12], etc. The remaining two aspects are relatively novel, so we won't repeat them here.

III. METHODS

Here we introduce Multithreading. This idea comes from an article about Multiple Immutable Memtables[9]. The article mentioned that since there is only one immutable Memtable structure in memory when the immutable Memtable flushes data to the disk, it will block the thread that the Memtable writes to it. If multiple immutable Memtables are created, this kind of problem will not occur, and the blockage caused by a single module is well solved here. Through our research, we found that when the immutable Memtable is filled and then flushed to the disk, it is a thread that is responsible for it. This results in the fact that even if there are more immutable Memtables, it can only queue up for flush data. We will provide a new one here. The idea is to open multiple threads to concurrently flush data so that it will be written to the disk at a high speed.

Fig 3 shows the most primitive state of LevelDB. There is only one Memtable and immutable Memtable component in memory, and the immutable Memtable is single-threaded when flushing data to the disk. Although this method can ensure data consistency, it also sacrifices More performance, and it is not friendly to the current multi-core CPU multi-threading, cannot play its full performance.

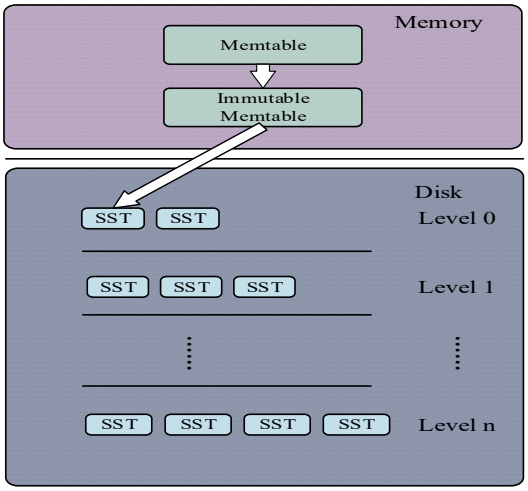


Figure 3:LevelDB Single Thread Mode

Fig 4 shows the optimization scheme of multiple immutable Memtable components and multithreading. It can be seen that multiple immutable Memtables can better optimize thread blocking, and multi-threaded flushing to disk can also optimize the long-tail delay problem.

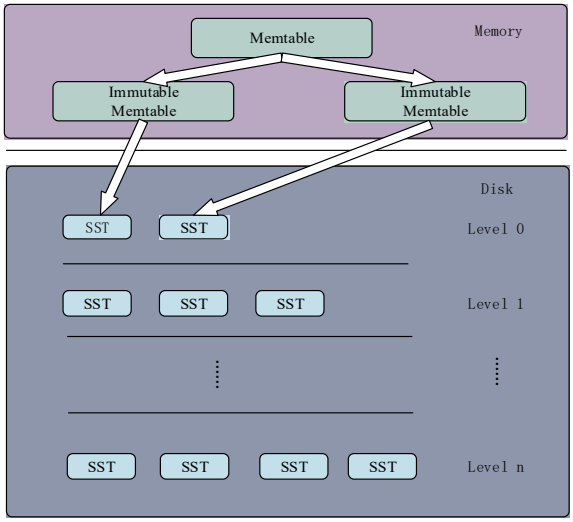


Figure 4:LevelDB Multithread mode

IV. EVALUATION

Experiment Setup. We conduct experiments on one machine running the Ubuntu 20.04 final with kernel 5.11.0. The machine includes two-sockets *Intel® Xeon®* Silver 4210(40 cores, 2.20GHz, 20MB L2 cache, 27.5MB L3 cache). The machine has 126GB of DRAM and nineteen 600GB disks. The file system used is ext4.

We tested fillseq, fillrandom, readseq, readreverse, overwrite these indicators respectively. The key size we use is 16 bytes, the value size is 100 bytes, and the workloads used are 1,000 and 10,000 respectively.

Results with benchmarks

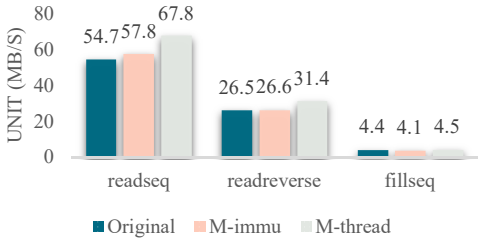


Figure 5:The comparative results of original, M-immu and M-thread under 1,000 Key/Value Workloads

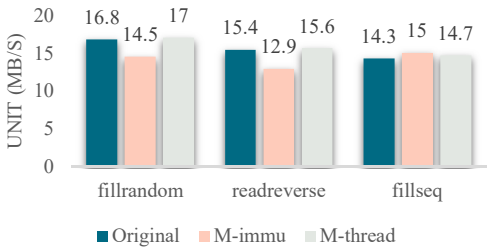


Figure 6:The comparative results of original, M-immu and M-thread under 10,000 Key/Value Workloads

Fig 5 shows the comparison of the three methods under 1000 key-value workloads. It clearly shows that the multi-threading method proposed in the article is better than the original method and the multi-immutable method. In readseq, readreverse, and fillseq, the improvement is 24%, 18%, and 2% respectively compared with the original method.

Fig 6 shows a comparison of the three methods under 10000 key-value workloads. It clearly shows that the multi-threading method proposed in the article is better than the original method and the multi-immutable method. In fillrandom, readreverse and fillseq, the improvement is 7%, 4%, and 3% respectively compared with the original method.

V. CONCLUSION

In this article, we introduced Multithreading, which is a new solution to improve reading and writing performance. It aims to increase the number of threads that interact with data from the memory to the disk to increase the rate of writing and reading data. And we have shown good results in the comprehensive workload experiment, which proves that this scheme is feasible.

ACKNOWLEDGMENT

This work is supported by The National Natural Science Foundation of China under Grant No.61762075.It is also supported by The Provincial Nature Science Foundation of Qinghai under Grant No.2020-ZJ-926. Ping Xie is the corresponding author of this paper.

REFERENCES

[1] O'Neil, P., Cheng, E., Gawlick, D., & O'Neil, E. (1996). The log-structured merge-tree (LSM-tree). Acta Informatica, 33(4), 351-385.

[2] LevelDB. 2005. A fast and lightweight key/value database library by Google. https://github.com/google/LevelDB.

[3] Raju, P., Kadekodi, R., Chidambaram, V., & Abraham, I. (2017, October). Pebblesdb: Building key-value stores using fragmented log-structured merge trees. In Proceedings of the 26th Symposium on Operating Systems Principles (pp. 497-514).

[4] Pan, F., Yue, Y., & Xiong, J. (2017). dCompaction: Delayed compaction for the LSM-tree. International Journal of Parallel Programming, 45(6), 1310-1325.

[5] Mei, F., Cao, Q., Jiang, H., & Li, J. (2018, October). SifrDB: A unified solution for write-optimized key-value stores in large datacenter. In Proceedings of the ACM Symposium on Cloud Computing (pp. 477-489)

[6] Balmau, O., Didona, D., Guerraoui, R., Zwaenepoel, W., Yuan, H., Arora, A., ... & Konka, P. (2017). {TRIAD}: Creating Synergies Between Memory, Disk and Log in Log Structured Key-Value Stores. In 2017 {USENIX} Annual Technical Conference ({USENIX}{ATC} 17) (pp. 363-375)

[7] Balmau, O., Dinu, F., Zwaenepoel, W., Gupta, K., Chandhiramoorthi, R., & Didona, D. (2019). {SILK}: Preventing Latency Spikes in Log-Structured Merge Key-Value Stores. In 2019 {USENIX} Annual Technical Conference ({USENIX}{ATC} 19) (pp. 753-766).

[8] Yao, T., Wan, J., Huang, P., He, X., Wu, F., & Xie, C. (2017). Building efficient key-value stores via a lightweight compaction tree. ACM Transactions on Storage (TOS), 13(4), 1-28.

[9] Gao, Y., Xie, P., Hua, W., Lv, M., & Li, P. (2021, August). Improve the Performance of LSM-Tree Based Key-Value via Multiple Immutable MemTables. In 2021 IEEE 12th International Conference on Software Engineering and Service Science (ICSESS) (pp. 223-227). IEEE.

[10] Sears, R., & Ramakrishnan, R. (2012, May). bLSM: a general purpose log structured merge tree. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (pp. 217-228).

[11] Lu, L., Pillai, T. S., Gopalakrishnan, H., Arpaci-Dusseau, A. C., & Arpaci-Dusseau, R. H. (2017). Wisckey: Separating keys from values in ssd-conscious storage. ACM Transactions on Storage (TOS), 13(1), 1-28.

[12] Ren, K., Zheng, Q., Arulraj, J., & Gibson, G. (2017). SlimDB: A space-efficient key-value storage engine for semi-sorted data. Proceedings of the VLDB Endowment, 10(13), 2037-2048.

Quantile Regression Method of House Price Index for Newly Building Condominiums in China

Ai Lirong
Wuhan Institute of Shipbuilding Technology
Wuhan 430050, China
e-mail: 3247438@qq.com

He Saiqi, Jin Shengping
School of Science
Wuhan University of Technology
Wuhan 430070, China
e-mail: 1410374478@qq.com
spjin@whut.edu.cn

***Abstract*—The house price index plays a very important role in the real estate economy due to it is an indicator of real estate price changing. Aiming at the compilation model of China's new ordinary residential housing price index based on the matching set of upright adjacent floors, the parameter estimation method is improved, and the ordinary least square regression (OLS) is replaced by quantile regression to eliminate the extreme sensitivity of least square regression to outliers. This paper puts forward the quantitative index to evaluate the advantages and disadvantages of different methods to compile house price index, and makes an empirical analysis of the above ideas based on the loan data of commercial banks. The conclusions we got are more consistent with the actual results, more stable, and in line with the requirements of the third generation of housing price index compilation methods.**

Keywords-house price index; set of upright adjacent floors; sample matching method; quantile regression

I. INTRODUCTION

The house price index is formulated through a certain number of real estate samples and a set of compilation methods based on statistical index theory. It is a relative value used to describe the magnitude and direction of changes in house prices and market prices in a certain area. Scientifically compiling house price index is the basic work of monitoring real estate price changes, and it has important theoretical and practical significance.

The compilation method of the house price index has undergone the first generation methods represented by the median and simple weighted average method, and the second generation methods represented by the sample matching method and the Laspeyres weighting method. The third generation method based on mathematical statistical analysis technology which uses the quality-controlled method has gradually become the development direction of the international housing price index compilation. At present, the two mainstream methods of quality adjustment are the repeat sales model and the hedonic model.

The National Statistics Bureau in China issued the residential sales price indices every month, which divides housing types into two major categories: newly building housing and second-hand housing. As shown in Fig. 1, new building housing is divided into residential and nonresidential housing, residential housing is divided into commercial housing and economic and functional housing, and commercial housing is divided into condominiums and high grade housing. Because of the history reason and house's quanlity, it is fitted to the situation in China at present that newly building housing and second-hand housing are treated and processed separative.

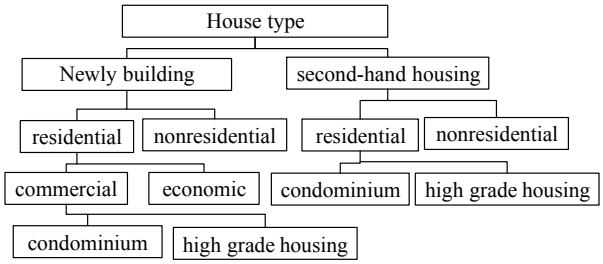


Figure 1. The categories of house type.

Currently, newly building condominiums of most cities in China account for the mainstream housing sales. The state has controlled strictly or prohibited the construction scale of high grade housing. Due to the economic and functional housings which are sold to the lower incomers, the prices are contored by the local goverments and not determined by the market. The prices data from the economic and functional housings should not be used to compile house price index. Indeed the National Statistics Bureau in China does not use the economic and functional housings to issue the residential sales price indices from January, 2018.

On the other hand, the western scholars have established method to compile the house price index for single houses[1]. Although there are condominiums price index such as in Manhattan, New York City, Singapor, Tokyo, these condominiums price index are compiled by using the methods for single houses and the price data for condominiums. They do not make use of the characteristics of the condominiums. So the theory and model to compile the condominiums price index have not been founded[2].

Therefore, it has an important theoretical and practical significance to study the price index compilation method of newly building condominiums in China. We hope that it is easy to achieve breakthrough from the theoretical aspect.

As for the compilation of the housing price index of newly building condominiums in China, most scholars focus on the hedonic method. Since the hedonic method requires the collection of data reflecting more than a dozen or even dozens of characteristic variables of each property, and the characteristic variables of houses may vary greatly in different cities or in different developing stages of the same city. In addition, some characteristic variables are

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00047

difficult to quantify. Therefore, the hedonic method has great limitations both in theory and in practice.

Chinese real estate has its own distinctive features. The newly commercial buildings are all high-rise structures, and most of them are developed in plots of complex. The external neighborhood characteristics of a same real estate or a same complex are basically the same. Intuitively speaking, the structure of newly commercial buildings in China is relatively 'single' or 'homogeneous', which is the core idea of the third generation method of house price index on homogeneity and comparability.

Some progress has been made in the research of the third generation method for compiling the house price index of newly building condominiums in China. Dr. Y. Xu of Xiamen University proposed the repeat-sales-like rule, which skillfully applied the repeat sales model to the new constructed housing market. Guo X. Y., Zheng S. Q., D. Geltner and Liu H. Y. of Tsinghua University proposed the pseudo repeat sales model[3].

However, the team of Xiamen University matched floors based on the property price and did not consider the impact of different orientations on property price, so its matching method may introduce new errors. The team of Tsinghua University used the 'single interval' principle to match, namely, after all samples are sorted by time, only the samples on adjacent time can be matched, which may cause errors due to the large difference in the distribution of properties in two adjacent periods.

McMillen[4] and Deng[5] used propensity scores and the team of Tsinghua University used a method similar to the hedonic method to construct matching pairs. Although the method is general, both of them have the same problem as hedonic price method, i.e. the selection bias of characteristic variables, which results in omitted characteristic variables in practical work and makes the method invalid or inaccurate.

Combined with the characteristics of the newly building condominiums in China, this paper studies the quantile solution method based on our previous work of compiling the house price index model of the newly building condominiums based on the set of upright adjacent floors, and puts forward the evaluation index which is used to compare and analyze the advantages and disadvantages of different methods[6-8].

II. THE MATCHING MODEL TO COMPILE THE HOUSE PRICE INDEX OF NEWLY BUILDING CONDOMINIUMS BASED ON THE SET OF UPRIGHT ADJACENT FLOORS

Urban residents in China are very sensitive to the orientation or location of house properties. For example, the prices of condominiums in the east side are generally much higher than those in the west side of a same building, and prices vary greatly on different floors. Based on the concept of the set of upright adjacent floors which we have put forward, we can deal with the problem that house prices are closely related to orientation or other neighborhood characteristics in the compilation of house price index.

A. The Semilinear Nonparametric Model of House Price Index

The main characteristics of commercial housing are structural characteristics and neighborhood characteristics. The structural characteristics include the floor, the area, the orientation of the house, the total floor of the building and so on. Neighborhood characteristics include factors such as the city area where it is located, and whether the surrounding public facilities are complete. Except for the floor, the other structural characteristics and neighbourhood characteristics of the same set of upright adjacent floors are the same. Table I is the main variables and their descriptions that will be used in the following.

TABLE I. THE MAIN VARIABLES AND THEIR DESCRIPTIONS

Variables	Description
Y_i, y_i	Y_i : Unit sales price of property i , unit: yuan; $y_i = \log(Y_i)$
fl_i, tf_i	fl_i : The floor where property i is located tf_i : The height of the building where property i is located
z_i	Standard height: $z_i = fl_i / tf_i$
t_i	The date (month) corresponding to the sale time of property i
tp_i	Type of building in which the property i is located, multi-storey buildings: $tp_i = 0$, small high-rise buildings: $tp_i = 1$, high-rise buildings: $tp_i = 2$
O_i	A set of other characteristics of property i

The semilinear nonparametric model of the logarithm of the house price is:

$$y_i = \beta_{t_i} + \alpha_1 z_i + \alpha_2 z_i^2 + \gamma_1 tp_i z_i + \gamma_2 tp_i z_i^2 + f(O_i) + \varepsilon_i. \quad (1)$$

Where,

1) β_{t_i} is the fixed effect of time t_i , which is the logarithm of the house price index, i.e. $\beta = (\beta_0, \beta_1, \cdots, \beta_T)^T$ denotes the logarithm of the house price index at all times, and β_0 is the logarithm of the house price index in the base period, let $\beta_0 = 0$.

2) α_1, α_2 are the fixed effect of the standard height and the square of the standard height of the floor, respectively. γ_1, γ_2 are the interaction effect of the standard height and its square with the type of buildings, respectively.

3) $f(O_i)$ denotes the influence of neighbourhood characteristics and other unobserved characteristics of the property i on housing prices.

4) ε_i denotes independent and $N(0, \sigma^2)$ -distributed error terms.

B. The Matching Model of House Price Index Based on Set of Upright Adjacent Floors

Assuming that commercial houses i and j belong to the same set of upright adjacent floors, then $O_i = O_j$, $tp_i = tp_j$, $tf_i = tf_j$. Applying within-pair first differencing based on model (1) will obtain (2):

$$y_i - y_j = \beta_{t_i} - \beta_{t_j} + \alpha_1 (z_i - z_j) + \alpha_2 (z_i^2 - z_j^2) + \gamma_1 tp_i (z_i - z_j) + \gamma_2 tp_i (z_i^2 - z_j^2) + \varepsilon_i - \varepsilon_j. \quad (2)$$

Equation (2) no longer contains function $f(O_i)$, so it does not need to collect various characteristic data of the real estate, and it does not required to consider the form of characteristic function, which solves many problems brought by the hedonic method.

For the distribution of $\varepsilon_i - \varepsilon_j$, we previously assumed that it was independent and identical distribution and had normal distribution, and used OLS to estimate the parameters in (2). This article studies the general situation and uses quantile regression to calculate.

III. QUANTILE REGRESSION OF MATCHED MODEL

The OLS regression method of (2) is based on the mean of the price, while the quantile regression estimates the parameters of any quantile of the price, such as selecting the median, 10% quantile, 90% quantile and so on. In the housing price index preparation, quantile regression is typically applied to the hedonic model, which can examine the changes in the degree of influence of each characteristic variable on the housing price at each quantile, and analyze the main factors for the increase or decrease of housing price. McMillen and Thorsnes(2006) applied quantile regression to the estimation of house price index by repeat sales method, to some extent, to suppress the influence of possible housing renovation or decoration on the overestimation of the housing price index in two transactions[9,10].

The quantile of regression is defined as follows:

$$Q_\tau(y \mid x) = \inf\{y : F(y \mid x) \geq \tau\}.$$

Where $F(y \mid x)$ is the conditional distribution function.

If the quantile regression function is a linear function $Q_\tau(y \mid x) = x^{\mathrm{T}} \beta_\tau$, then the linear quantile regression model is:

$$y_i = x_i^{\mathrm{T}} \beta_\tau + \varepsilon_i, i = 1, \cdots, n. \quad (3)$$

Where any τ -th quantile of ε_i is 0. Then, for any $0 < \tau < 1$, the quantile regression problem can be transformed to solve the following problem:

$$\hat{\beta}_\tau = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau (y_i - x_i^{\mathrm{T}} \beta).$$

Where $\rho_\tau(Z) = Z(\tau - \mathrm{I}(Z < 0))$. When $\tau = 1/2$, it is equivalent to minimize the absolute value of the error, which is the so-called median regression or L_1 norm regression.

The main idea of quantile regression is to modify the objective function of OLS regression, which is very sensitive to outliers. The objective function of quantile regression in model (2) is:

$$\min \sum \rho_\tau \begin{pmatrix} y_i - y_j - \beta_{t_i} + \beta_{t_j} - \alpha_1 (z_i - z_j) \\ -\alpha_2 (z_i^2 - z_j^2) - \gamma_1 tp_i (z_i - z_j) \\ -\gamma_2 tp_i (z_i^2 - z_j^2) \end{pmatrix}. \quad (4)$$

Equation (4) can be solved by simplex method or interior point method of linear programming, or by functions of some software packages, such as rq() in the package 'quant' of R. The determination of the quantile should be based on the sample data and the reasonable range of the housing price change, through a certain number of indicators to evaluate and finally determine the value of the quantile.

IV. EMPIRICAL CALCULATION AND RESULT ANALYSIS

A. Data Source and Calculation Results

The Wuhan Branch of the People's Bank of China collected the mortgage loan data from commercial banks of 83 newly sold commercial houses in Xiangyang from January to December 2012. The transaction price is real and reliable, which can reflect the market situation of new building condominiums in Xiangyang.

After eliminating incomplete information, 6,354 valid samples were obtained. According to (2), the matching process based on upright adjacent floors was carried out, and 4,579 matching pairs were obtained, accounting for 4579/6354=72.1% of the original data.

Using the OLS calculation parameters, the house price index corresponding to the row of 'OLS' in Table II is obtained, where the base index of January is 100. Then, use the quantile regression method to estimate parameters in the model (2) with the data set of all 4,579 matching pairs, and let $\tau = 0.3, 0.5, 0.8$ respectively. The calculation results are shown in the corresponding rows of 'Quant0.3, Quant0.5, Quant0.8' in Table II.

TABLE II. ESTIMATE RESULTS OF OLS METHOD AND QUANTILE REGRESSION

Month / Methods	1	2	3	4
OLS	100.000	100.380	100.488	101.023
Quant0.3	100.000	100.206	100.459	100.375
Quant0.5	**100.000**	**100.016**	**100.063**	**100.157**
Quant0.8	100.000	99.952	100.359	100.639
Month / Methods	**5**	**6**	**7**	**8**
OLS	100.834	101.260	101.632	102.212
Quant0.3	100.633	100.416	100.687	101.116
Quant0.5	**100.427**	**100.323**	**100.442**	**100.701**
Quant0.8	100.449	100.328	100.486	100.850
Month / Methods	**9**	**10**	**11**	**12**
OLS	102.666	103.097	102.833	102.639
Quant0.3	101.444	101.349	101.369	101.313
Quant0.5	**100.794**	**100.806**	**100.810**	**100.810**
Quant0.8	101.145	101.557	101.370	101.459

The housing price index in Table II is drawn as a corresponding line chart, as shown in Fig. 2.

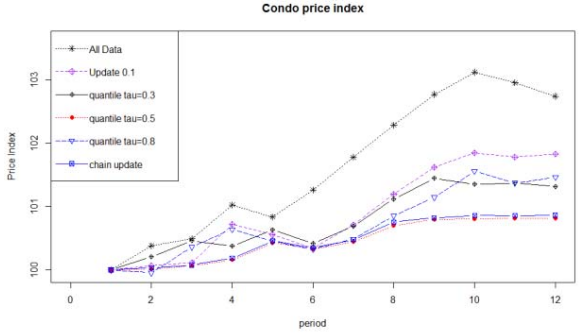


Figure 2. The OLS Index and the Quantile Indices of Xiangyang.

It can be seen from Fig. 2 that the OLS regression and the quantile regression with a quantile far away from 0.5 overestimate the house price index, and the quantile regression with a quantile of 0.5 suppresses matching pairs with too large errors, so the trend of corresponding house price index is relatively flat.

B. The Calculation Results Analysis

In order to compare the advantages and disadvantages of various methods, the following tests are carried out. Randomly select 90% of 4,579 matching pairs as the training set, and the remaining 10% as the test set.

Based on the training set, use OLS method and quantile regression to estimate the parameters in model (2). The estimated parameters are applied to each matching pair (i, j) in the test set, and the predicted value $\hat{y}_i$ of one of the commercial houses can be obtained. This predicted value and its actual price. The difference between the predicted value $\exp(\hat{y}_i) = \hat{Y}_i$ and the true value can be used to test the advantages and disadvantages of the calculation method of the model.

Define the root mean square error($RMSE$) as shown in (5):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}. \quad (5)$$

Where n is the number of matching pairs in the test set, and $Y_i, \hat{Y}_i$ denote the true and predicted house prices of commercial housing i respectively.

Randomly select 457 matching pairs from all data sets as the test set, and the rest as the training set. After a random segmentation, we can get the corresponding $RMSE$ after using OLS and quantile regression to estimate the parameters of (2). Repeat the experiment for 500 times, then calculate the mean value and standard deviation of 500 corresponding $RMSE$ to the two methods respectively, and obtain the corresponding Mean value(Mean(RMSE)), minimum value(Min(RMSE)) and standard deviation(Sd(RMSE)), as shown in Table III:

TABLE III. THE VALUE OF EVALUATION INDEX OF VARIOUS METHODS

Indicators / Methods	Mean Value Mean(RMSE)	Minimum Value Min(RMSE)	Standard Deviation Sd(RMSE)
OLS	292.590	188.140	53.406
Quant0.3	292.011	183.649	52.665
Quant0.5	**291.786**	**184.701**	**52.553**
Quant0.8	293.762	189.771	52.754

Based on the mean, minimum and standard deviation of $RMSE$, the smaller values, the better. Compared with quantile 0.5, the mean, minimum and standard deviation of OLS, quantile 0.3, 0.8 and other quantiles (not listed in Table 3) are relatively larger (except for the minimum value corresponding to quantile 0.3). Thus, the quantile 0.5 is a reasonable choice and an expected conclusion for the calculation of the housing price index.

In Fig. 2, the housing price index corresponding to 'Quant0.5' is smaller than the house price index corresponding to 'All Data' and 'Update0.1', indicating that OLS method greatly overestimates the house price index, mainly because the OLS method is highly sensitive to data with large errors.

V. CONCLUSIONS AND SUGGESTIONS

According to the matching model based on the set of upright adjacent floors proposed by the author's team, this paper will use the quantile regression to study the method of solving the parameters, and put forward quantitative indicators to evaluate the advantages and disadvantages of different compilation methods. The calculation example shows that the quantile regression with quantile 0.5 can reflect the housing price market more stably and truly.

In this paper, the matching model and method based on the set of upright adjacent floors are improved appropriately, which can also be applied to the compilation of the housing price index of the second-hand house. Also, they provide an early theoretical preparation for the compilation of a unified general commercial housing price index when the newly building house and second-hand house market are merged together.

REFERENCES

[1] K. E. Case and R. J. Shiller, "The efficiency of the market for single-family homes," The American Economic Review, vol. 79, Feb. 1988, pp. 125-137, doi:10.3386/w2506.

[2] M. B. David, "S&P CoreLogic Case-Shiller Home Price Indices Methodology," Copyright © 2016 S&P Dow Jones Indices LLC, https://us.spindices.com/index-family/real-estate/sp-corelogic-case-shiller.

[3] Guo X. Y., Zheng S. Q., D. Geltner and Liu H. Y., "A new approach for constructing home price indices: The pseudo repeat sales model and its application in China," Journal of Housing Economics, vol. 25, Sep. 2014, pp. 20-38, doi:10.1016/j.jhe.2014.01.005.

[4] D. P. McMillen, "Repeat sales as a matching estimator," Real Estate Economics, vol. 40, Sep. 2012, pp. 745-773, doi:10.1111/j.1540-6229.2012.00343.x.

[5] Deng Y. , D. P. McMillen and T. F. Sing, "Private residential price indices in Singapore: a matching approach," Regional Science and Urban Economics, vol. 42, May 2012, pp. 485-494, doi:10.1016/j.regsciurbeco.2011.06.004.

[6] Jin S. P., Zeng X. and Li Q., "Research on the model to compile the house price index for newly building condominiums in China," Wuhan Finance, vol. 6, 2017, pp.49-53, doi:CNKI:SUN:YHQY.0.2017-06-012.

[7] Jin W. H. and Jin S. P., "The Virtual repeat sale model for the house price index for new building in China," Applied Mathematics, vol. 5, Dec. 2014 , pp. 3431-3436, doi: 10.4236/am.2014.521320.

[8] Dong Y. L., Jin S. P. and Chen J. Q., "Semiparametric matching model for compiling price index of newly-built ordinary houses," Statistics and Decision, vol. 36, Aug. 2020, pp. 41-44, doi:10.13546/j.cnki.tjyjc.2020.16.009.

[9] D. P. McMillen and T. Paul, "Housing renovations and the quantile repeat‐sales price index," Real Estate Economics, vol. 34, Nov. 2006, pp. 567-584, doi:10.1111/j.1540-6229.2006.00179.x.

[10] Zhang L. and Yi Y., "Quantile condominium price indices in Beijing," Regional Science and Urban Economics, vol. 63, Mar. 2017, pp. 85-96, doi:10.1016/j.regsciurbeco.2017.01.002.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

A hybrid ARIMA–ANN model for Internet Food safety risks

Wei Wang
School Of Artificial Intelligence and Computer Science,Jiangnan University
WuXi ,China
E-mail:321810255@qq.com

Jun Sun
School Of Artificial Intelligence and Computer Science,Jiangnan University
WuXi ,China
E-mail:junsun@jiangnan.edu.cn

***Abstract*—As Internet technology becomes more and more mature, Internet food sales have become more and more problematic. problems from raw materials to food processing to food distribution are not easily regulated because of online sales. Firstly , crawl the take-out sales data of cities across the country, based on the data the time series prediction model is established to realize the prediction of business and regional risk values, From this obtains the merchant early warning as well as reminds the consumer's function. According to the data analysis, the ARIMA+ANN model is used for short-term prediction. Then the data analysis results of different regions are compared, and the data analysis of both sides of time and space is conducted, The value at risk ranking allows consumers to make a clear comparison of effects, In this way, we can realize the real and effective supervision of store owners from the economic level, realize the real and effective supervision of Internet food sales, and ensure food safety.**

***Keywords-component;* Internet Food safety ; Data acquisition and cleaning; spatiotemporal series; Arima; ANN**

I. INTRODUCTION

With Food safety incidents have occurred frequently In recent years, causing all walks of life to pay attention to food safety issues. With the gradual penetration of Internet thinking into traditional industries, the Internet will play an increasing role in the field of food safety [1], 2019, The number of Internet food delivery users has reached 421 million in china, The increase of 15.16 million compared to 2018, accounting for 49.3% of the total number of Internet users. The number of takeaway users has reached 417 million, Compared with the increase of 20.37 million in 2018, accounting for 49.3% of mobile Internet users, various drug takeout incidents emerge in endlessly, so Internet food safety is an urgent problem to be solved. Through the collection, screening and analysis of consumers' evaluation of various businesses, this project can have a specific and digital understanding of the food safety and health problems of businesses in this period, and then make a prediction of food safety of businesses through these data, By publicizing the results to the public[2], the public can have an intuitive understanding and comparison of food after obtaining information through various media.

From a theoretical point of view: according to the project planning, first of all, according to the "The PRC Food Safety Law" and "Food Safety Supervision And Management Measures of Online Catering Service", the indicators of food safety violations are determined [3], and the data of Internet food are classified reasonably, so as to have a more accurate cognition of each store.

In fact, the results of the experiment have guiding significance for both businesses and consumers. Businesses can grasp the safety status of their own stores according to the relevant scores and the recent scoring trend, and timely pay attention to adjust the food safety problems in the stores [4]. For consumers, the level of food safety score can be used as an important indicator of choosing consumption which is very important for food safety The prediction of food safety can also give consumers a relatively clear understanding of the recent food safety trend of a store [5], and the regional analysis also has a certain reference value for government supervision.

Big data technology has been applied to various fields of food safety in China, and scholars have been committed to establishing a huge food safety database and a visual model of food safety analysis [6].

Combined with computer technology, there are many applications in the direction of food safety, such as detection technology in the field of food safety, including infrared spectrum, hyperspectral image, computer vision, artificial olfaction, biochip and other detection technologies [7]. In the aspect of food safety information traceability, there is also a food information traceability system based on radio frequency identification (RFID), which can send and receive chip information, so as to obtain relevant food information and establish a traceability system [8], so as to detect the purpose of food safety on the other hand

There are some loopholes in China's current food safety legal regulation in the legal system, law enforcement and other aspects. This experimental model can play an early warning role to a certain extent [9].

The rest of this paper is arranged as follows: the first chapter is the data source and pretreatment, the second chapter is the construction of Internet food risk prediction model, the third chapter is the experimental results and analysis, the fourth chapter is the comparison of arima-ann, ARIMA, LSTM model effect, and the last chapter is the conclusion and reference literature.

II. DATA SOURCE AND PRETREATMENT

This chapter mainly introduces the data source and data preprocessing of this article. First, the analytic hierarchy process is used to obtain the weights, and then the Bert model analysis is used to obtain the food safety risk value, and then the data preprocessing includes time series stationarity detection and unit root detection.

A. Date Source

Use the Scrapy framework to obtain takeaway information in 46 major cities across the country, totaling approximately 230,000 takeaway stores and 38 million

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00048

review information, and filter out stores with too small data. The final data used in the experiment is about 200,000 stores and 3,600. Million comments.

B. Data Preprocessing

First, the time series stationarity test is performed to test the stationarity of the sequence, and the difference is when the sequence is not stable. Since the data is stable mainly by subjective experience, it is necessary to perform unit root test to confirm whether the data is stable after the difference, and there is an objective judgment.

1)Using the analytic hierarchy process [10-13] to obtain the weights of 16 categories, the steps are as follows: invite a number of people familiar with the situation as scorers, ask them to fill in the designed questionnaire, then the judgment matrix table is obtained, and the characteristic vector of the judgment matrix is calculated . The steps are: Calculate the product of each row of the judgment matrix:

$$M_i = \prod_{j=1}^{n} b_{ij} \quad (2)$$

Calculate Mi=n times square root Vi:

$$V_i = \sqrt[n]{M_i} \quad (3)$$

Normalize Wi: $W_i = V_i \sum V_i$

$$W_i = v_i \sum v_i \quad (4)$$

It is obtained that Wi=(W1, W2, W3,..., Wn) is the eigenvector of the judgment matrix, that is, the corresponding weights are shown in Table 1.

Table 1 Indicator weight map

index	***ranking***
C1Use perverted raw materials	0.309975
C4Merchants operate in an unsothy environment	0.1382
C3There is a mix of foreign objects in the food	0.109617
C6Provide unqualified cutlery	0.084708
C5Use nonconforming packaging materials	0.061817
C2Food processing procedures are improper	0.052617
C9The sanitary condition of the distribution container	0.051267
C11Sell out-of-date food	0.046342
C8Special food quality	0.035192
C7The packaging is not standardized	0.032475
C10The product label is not qualified	0.016517
C15Merchants provide false information	0.015858
C16Merchants swipe orders and fake comments	0.01265
C13Merchants do not process or delay handling complaints to report incidents	0.012458
C12Inso- and offline inse and out of line	0.01025
C14The merchant does not provide proof of sale	0.010067

The data is labeled and substituted into the Bert[14-16] training model. The Bert model is an NLP model developed by the Google AI team in 2018. The main process is as follows: each word (token) in the comment is sent to the embedding layer, plus Self-attention mechanism, and then training in the network, and finally get the probability of 16 classifications (this experiment uses the Bert model as part of the data preprocessing, so it is briefly described), and then multiply by the index weight obtained by the analytic hierarchy process, then A score can be obtained for each review, all review information is summarized, and the weighted average is finally used to obtain the risk value of the merchant in each time period through the model.

2) Time Series Stationarity Detection

First judge the stability of the data, and judge that the group of data has no obvious upward and downward trend, and there is a continuing trend, that is, whether there is a finite fluctuation around a certain value. If the data is not stable, the data needs to be differentiated. Until the data is stable

3) Unit Root Detection

Then objectively use the unit root test to test whether the time series is stable after the score, and whether there is a unit root in the risk value to test whether the data is stable. If it does not exist, it is stable. Otherwise, it is not stable. The formula is as follows:

$$\Delta yt = (\rho - 1) yt - 1 + ut$$
$$= \delta yt - 1 + ut \quad (5)$$

The whole formula is a regression model, Δ is the first difference, t is the time, ρ is the coefficient, and ut is the error term. Testing whether there is a unit root is equivalent to testing whether = 0

4) After the processed data is stable, the data needs to be tested for white noise. If the data is white noise, it is a set of purely random numbers. The white noise monitoring in this experiment is divided into two parts: raw data white noise monitoring , Residual data white noise monitoring, using the Ljung-Box method to monitor white noise, the specific process is described in the experiment.

III. Internet Food Risk Prediction Model Construction

The risk prediction model is ARIMA+ANN[17-20] combined prediction. The main process is ARIMA[21-24]. First make predictions, and then put the data residual values into the ANN model. The two model results are combined to obtain the final prediction value.

A. ARIMA Model

The ARIMA model is also known as the differential integrated moving average regression model. First of all, the premise of this model is that the data must be a stationary sequence and have a continuation trend. First, the sequence to be processed is subjected to d-order difference until the data is stable. The sequence model is ARIMA(p,d,q), p is the order of the autoregressive model, and q is the moving average order. P is solved by plotting the autocorrelation coefficient ACF and partial autocorrelation coefficient PACF of the sequence, which are specifically determined by analyzing the graph.The structure of the ARIMA time series model is shown in Figure 1:

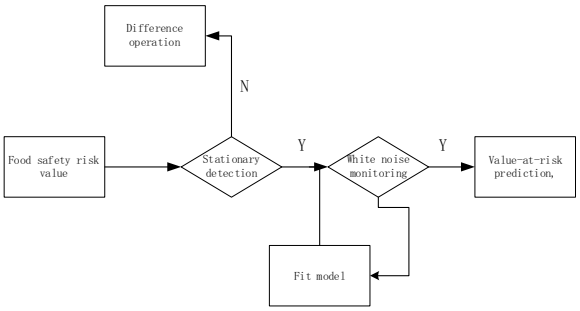


Figure 1 ARIMA time series model structure

B. *ARIMA-ANN model*

Biological neural networks are connected by a large number of biological neurons. Similarly, ANN is also composed of multiple neuron models connected according to certain rules.

Use Keras to build a neural network, mainly including input layer, hidden layer, and output layer. Through the training of sample data, the network weights and thresholds are constantly modified to make the error function fall along the negative gradient direction, and the hidden layer can be one or more layers.

In this article, the ANN neural network is used to predict the residual data generated by the ARIMA model. Then the residual prediction results and the ARIMA prediction results are fused, the hidden layer uses the tanh transfer function, and the output layer uses a linear function.

The residual prediction steps of the ANN network in this article are as follows:

(1) Determine the number of nodes according to the output results of the ARIMA model, and determine the input data and output data

(2) Determination of the number of hidden layer nodes: In fact, enough hidden layer nodes can fit any non-linear function, but the corresponding calculation amount will increase correspondingly, and the phenomenon of overfitting will affect the model Effect, on the other hand, too small number of nodes will also lead to poor model performance. In fact, the number of hidden nodes is related to the complexity of the actual problem, the number of inputs and outputs, and the expected error. The calculation formula for the selected hidden layer used in this article is:

$$L=\sqrt{\mathrm{m+n}}+a \qquad (5)$$

Among them, n is the number of nodes in the input layer, m is the number of nodes in the output layer, and a is a constant between [1,10] to determine the range of the number of nodes in the hidden layer.

The basic structure of the ARIMA-ANN fusion model is shown in Figure 2:

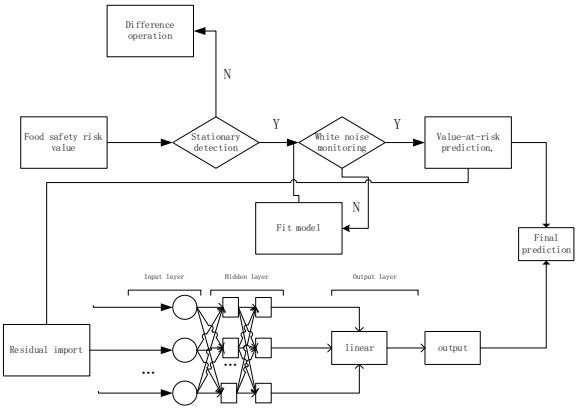


Figure 2 Mixed model diagram

The main process is as follows:

(1) After obtaining the merchant's risk value, the sequence is brought into the ARIMA model. First, the data stationarity is detected. If it is not stable, the difference is performed until it is stable, and then it is placed in the trained ARIMA model to predict the preliminary result y1.

(2) Make the difference between the ARIMA model prediction result and the true value to obtain the residual value, and import the residual value into the ANN,

Get the residual prediction value y2.

(3) Combine the ARIMA model result y1 with the residual prediction result y2 obtained by ANN to obtain y:

$$Y_{\mathrm{f}}=Y_1+Y_2 \qquad (6)$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

First, take out a store for model testing. Take the time period from July 2016 to July 2019 as an example, the interval period is one week, and there are a total of 145 data. As shown in Figure 3

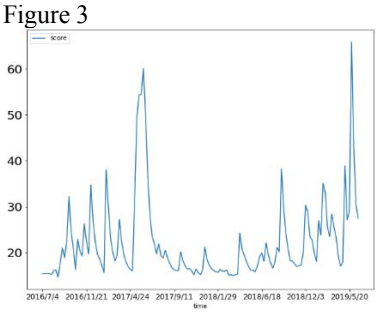


Figure 3 Data timing

The first 66% of the data is used as the training set, and the remaining 34% is used as the test set, that is, the first 95 pieces of data are used as the training set, and the last 50 pieces of data are used as the test set. Carry on the stationarity detection of the data, the difference detection effect is shown as in Fig. 4.

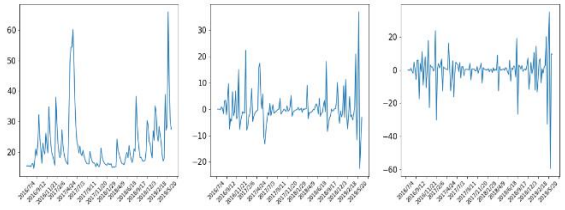


Figure 4 Differential effect

A. *Stationarity detection of data series*

Figure 4 shows the original data, the first-order difference, and the second-order difference. The time series

data is basically stable after the first-order difference. The ADF (unit root) detection of the first-order difference results is shown in Table 2.

The result shows that the data fluctuates relatively smoothly after the difference. The unit root test value adf is less than one percent, five percent, and 13 percent. The P-Value value is 4.6341*10-5, which is close to 0. , Data stationary series no longer need difference operation.

Table 2 ADF detection

adf	cValue			p 值
	1%	5%	10%	
-4.836	-3.477	-2.882	-2.577	0.00004

B. White noise detection of original data

The Ljung-Box method is used to test, and the Ljung-Box test is the LB test, which is a method to test the autocorrelation of the series in the time series analysis. The Q statistic of the LB test is:, where T is the sample size, m is an artificially selected number, and ρi is the autocorrelation coefficient of the i-order lag. The experimental results are shown in Table 3. The p value obtained from the original data is much less than 0.05, so the data is a stationary non-white noise sequence.

Table 3 Raw data white noise monitoring

Stat	P
80.45290731	2.97716222*10-19

C. ARIMA Model

First determine the three parameters p, d, and q required by the ARIMA model. The first difference has been determined above, so the value of d is 1.

Plot ACF (autocorrelation coefficient graph) and PACF (partial autocorrelation coefficient).

The autocorrelation coefficient compares an ordered sequence of random variables with itself. It reflects the correlation between the same sequence in different time series. In fact, the autocorrelation coefficient is not only the relationship between two elements, but also its influencing factors. There are all the elements between these two elements, and the partial autocorrelation coefficient removes the intermediate influencing factors, which is a strict correlation between the two elements. As shown in Figure 5

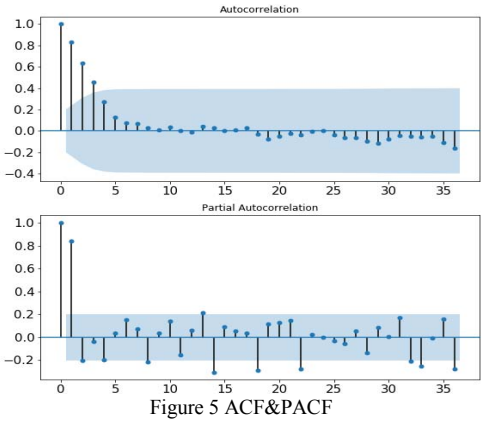


Figure 5 ACF&PACF

From the ACF and PACF diagrams, we can get the ACF diagram is censored, and the PACF diagram is trailing, so p and q are (1, 0), that is, we can try to fit the data sequence with the ARIMA (1, 1, 0) model.

However, considering the large number of stores and human identification will be accompanied by great subjectivity and historical experience, the optimal model identification method is used to calculate AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and select the smallest AIC, BIC values, and finally the ARIMA (1, 1, 0) model is used to fit the data sequence, and the summary information of the final model is shown in Table 4.

Table 4 ARMA Model Results

Dep. Variable:	y	No. Observations:	294
Model:	ARMA(1, 0)	Log Likelihood:	-988.256
Method:	css-mle	S.D. of innovations:	6.970
AIC:	1982.513	BIC:	1993.564
Sample:	0	HQIC:	1986.938

The model prediction results are shown in Figure 6:

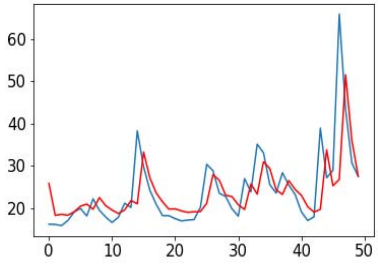


Figure 6 ARIMA_PREDICT

D. ARIMA-ANN model prediction

The ARIMA (1,0,1) model is used above to model the takeaway risk value sequence of a store in Wuxi from July 2016 to July 2019, fitting the linear information of the time series, the ARIMA prediction result is y1, and the ANN model prediction The result is y2, the final predicted result. The effect diagram of model prediction results is shown in Figure 7.

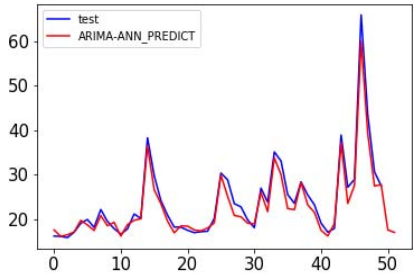


Figure 7 ARIMA-ANN model renderings

E. Comparison of the effects of ARIMA-ANN, ARIMA, and LSTM models

LSTM is a special form of RNN, mainly to avoid the problem of long-term dependence. It is solved by introducing a number of gates. The LSTM architecture has an input layer, a hidden layer and an output layer. The hidden layer is composed of memory cells. Each cell has three gates, namely: input gate, forget gate and output gate. Each gate can access the current input and previous output [25-27].

The comparison of the effects of ARIMA-ANN, ARIMA, and LSTM models is shown in Figure 8. In the experimental results shown in the figure, the blue line is the actual experimental value, the red is the predicted value of the ARIMA-ANN model, the green is the predicted value of the ARIMA model, and the yellow is the LSTM The prediction value of the model can be concluded that the

ARIMA model can predict the trend of food safety risks, but the prediction error is large. The prediction effect of the LSTM model is similar to the result of the ARIMA model, and there is also a prediction lag effect. In fact, time series analysis is being performed. It's normal to have a lag at the time. Generally, it is caused by insufficient information. In this article, it may be because the risk value time series are more complicated, and there are more data in the nonlinear time series. After the ARIMA-ANN model further processes the residuals generated by the ARIMA model, the effect achieved Better, the delay phenomenon is better solved in this data sequence, and the prediction error value is also small.

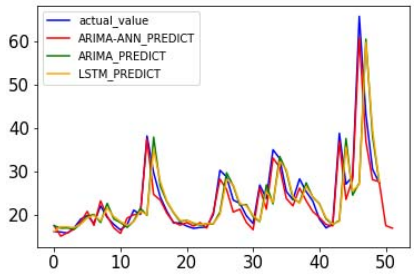


Figure 8 Comparison of prediction models

V. CONCLUSION

This paper uses the ARIMA+ANN model to realize the food safety risk value modeling, which greatly improves the lag of the ARIMA model and the LSTM model in this experiment. Through model testing, the ARIMA+ANN model has good predictive capabilities. The model can make short-term predictions of the merchant's risk value. It has certain guiding significance for government supervision, store self-recognition, and customer consumption. Spatio-temporal analysis The result forecast is also conducive to deepening the government's awareness of local food safety.

VI. REFERENCES

[1] Xu Wei,Zhang Zhipeng,Wang Hongxun,Yi Yang,Zhang Yanpeng. Optimization of monitoring network system for Eco safety on Internet of Things platform and environmental food supply chain[J]. Computer Communications,2020,151(C).

[2] Meng Chun,Sun Lin,Guo Xiaoni,Wu Miao,Wang Yuqi,Yang Lingping,Peng Bin. Development and Validation of a Questionnaire on Consumer Psychological Capital in Food Safety Social Co-governance [J]. Frontiers in Psychology,2021.

[3] Hernández San Juan Isabel. The Blockchain Technology and the Regulation of Traceability: The Digitization of Food Quality and Safety[J]. European Food and Feed Law Review,2020,15(6).

[4] Tsai Hsinyeh,Lee Yu Ping,Ruangkanjanases Athapol. Understanding the Effects of Antecedents on Continuance Intention to Gather Food Safety Information on Websites[J]. Frontiers in Psychology,2020.

[5] Francesco Meneguzzo,Federica Zabini. Agri-food and Forestry Sectors for Sustainable Development[M]:2021-02-26.

[6] Kharroubi Samer,Nasser Nivin A,ElHarakeh Marwa Diab,Sulaiman Abdallah Alhaj,Kassem Issmat I. First Nation-Wide Analysis of Food Safety and Acceptability Data in Lebanon.[J]. Foods (Basel, Switzerland),2020,9(11).

[7] Mojun Zou. Analysis of Food Safety Inspection Technology and Method[J]. International Journal of Education and Economics,2020,3(4).

[8] Drangert Jan-Olof. Correction to: Urban water and food security in this century and beyond: Resource-smart cities and residents[J]. Ambio,2021,50(3).

[9] By Thomas W. Hertel,Uris L.C. Baldos,Keith O. Fuglie. Trade in Technology: A Potential Solution to the Food Security Challenges of the 21st Century[J]. European Economic Review,2020.

[10] Annas Firdaus,Ediana Dina,Kurniawan Asep,Wandira Raju,Zakir Supratman. Decision Support System in Detrmination of Project Tender Winner Using the Analytical Hierarchy Process (AHP) Method[J]. Journal of Physics: Conference Series,2021,1779(1).

[11] Qing Wang,Song Liu,Congcong Li,Chao Yu,Yanxi Liu. AHP Based Evaluation Model of Energy Efficiency for Energy Internet[J]. E3S Web of Conferences,2021,233.

[12] Pham Ngoc Thach,Do Anh Duc,Nguyen Quang Vinh,Ta Van Loi,Dao Thi Thanh Binh,Ha Dieu Linh,Hoang Xuan Truong. Research on Knowledge Management Models at Universities Using Fuzzy Analytic Hierarchy Process (FAHP)[J]. Sustainability,2021,13(2).

[13] Olabanji Olayinka Mohammed,Mpofu Khumbulani. Appraisal of conceptual designs: coalescing Fuzzy Analytic Hierarchy Process (F-AHP) and Fuzzy Grey Relational Analysis (F-GRA)[J]. Results in Engineering,2020(prepublish).

[14] Wang Junshu,Zhang Guoming,Wang Wei,Zhang Ka,Sheng Yehua. Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT[J]. Journal of Cloud Computing,2021,10(1).

[15] [Engineering; Researchers at Chinese Academy of Sciences Target Engineering (Enhancing BERT Representation With Context-Aware Embedding for Aspect-Based Sentiment Analysis)[J]. Journal of Technology & Science,2020.

[16] Su Jing,Dai Qingyun,Guerin Frank,Zhou Mian. BERT-hLSTMs: BERT and hierarchical LSTMs for visual storytelling[J]. Computer Speech & Language,2021,67.

[17] Pavan Kumar Singh,Nitin Singh,Richa Negi. Wind Power Forecasting Using Hybrid ARIMA-ANN Technique[M].Springer Singapore:2019-03-31.

[18] Ümit Çavuş Büyükşahin,Şeyda Ertekin. Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition[J]. Neurocomputing,2019,361.

[19] C Narendra Babu,Pallaviram Sure. Partitioning and interpolation based hybrid ARIMA–ANN model for time series forecasting[J]. Sādhanā,2016,41(7).

[20] Ümit Çavuş Büyükşahin,Şeyda Ertekin. Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition[J]. Neurocomputing,2019,361.

[21]]Singh Ram Kumar,Rani Meenu,Bhagavathula Akshaya Srikanth,Sah Ranjit,Rodriguez-Morales Alfonso J,Kalita Himangshu,Nanda Chintan,Sharma Shashi,Sharma Yagya Datt,Rabaan Ali A,Rahmani Jamal,Kumar Pavan. Prediction of the COVID-19 Pandemic for the Top 15 Affected Countries: Advanced Autoregressive Integrated Moving Average (ARIMA) Model.[J].JMIR public health and surveillance,2020,6(2).

[22] .Periodic autoregressive models for time series with integrated seasonality[J]. Journal of Statistical Computation and Simulation,2021,91(4).

[23] Makala D,Li Z. Prediction of gold price with ARIMA and SVM[J]. Journal of Physics: Conference Series,2021,1767(1).

[24] COVID-19/SARS-CoV-2 News from Preprints; Brief Analysis of the ARIMA model on the COVID-19 in Italy (Published April 11, 2020)[J]. Medical Letter on the CDC & FDA,2020.

[25] Wei Li,Amin Kiaghadi,Clint Dawson High temporal resolution rainfall–runoff modeling[J],Neural Computing and Applications, 2020(prepublish), pp.1-18

[26] Mohamed Hosny,Minwei Zhu,Wenpeng Gao,Yili Fu. A novel deep LSTM network for artifacts detection in microelectrode recordings[J]. Biocybernetics and Biomedical Engineering,2020.

[27] Zhou Yutao,Wu Huayi,Cheng Hongquan,Qi Kunlun,Hu Kai,Kang Chaogui,Zheng Jie. Social graph convolutional LSTM for pedestrian trajectory prediction[J]. IET Intelligent Transport Systems,2021,15(3).

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Model checking the Efficiency of Blockchain-based Edge Computing Network

Kai Zheng, Xiang Yao,Zhe Zhang, Liyou Fang
Taihu University of Wuxi
Jiangsu Key Construction Laboratory of loT Application Technology
Wuxi , China
534347752@qq.com

Xin Huang
Xi'an Jiaotong-liverpool University
Suzhou , China
zg570419@126.com

***Abstract*—In last 10 years, the development of cloud computing was very fast. Cloud computing enables the users to obtain remote computing resources and brings convenience to them. However, the increasing information exchange makes the efficiency of cloud computing becomes a problem. In order to solve the problem, edge computing emerged. For edge computing, the edge devices can deal with a part of work load. In this way, the work load of cloud can be reduced and the efficiency of the whole network can be.enhanced. For some edge computing networks, blockchain is used to increase the security and information transparency of the whole network. However, how to further optimize the efficiency of the blockchain-based network is still a challenge. In this paper, discrete-time Markov chain (DTMC) model is built to evaluate the efficiency of the blockchain-based network. The main influence factors for the efficiency of it will be find, and the strategies to improve the efficiency of it will be find.**

Keywords- cloud computing; edge computing ;blockchain; DTMC;

I. INTRODUCTION

In last 10 years, cloud computing has developed very fast and it has applications in many occasions like e-commerce, power system, IoT system and so on[1-3]. Cloud computing also brings more powerful computing resources to users.

However, the fast increasing information exchange makes the efficiency of cloud computing becomes a challenge. In order to solve the problem, the concept of edge computing appears. For edge servers in edge computing, they can deal with a part of work load and usually they are located near the user side. Thus, the efficiency of edge computing is better than cloud computing. In recent years, the application of edge computing is increasing.

In some edge computing network, blockchain is used to enhance the security and information transparency of the whole network [4]. The fist application of blockchain is bitcoin[5]. But many countries still do not want to encourage bitcoin trading. However, as the basic technology of bitcoin, blockchain began to get people's attention. Many areas use blockchain technology[6]. Many governments like Chinese government are also interested in developing blockchain technology to enhance the information security and transparency.

However, how to further improve the efficiency of blockchain based edge computing network become a challenge. Some researchers also did some research in this areas. In [7], an edge computing accelerator has been proposed to improve the efficiency of edge computing network. For blockchain network, there are also some researchers focus on improve the efficiency of it. According to [8], a coordinated satellite-terrestrial network is proposed to improve the efficiency of the blockchain network.

However, few of the researchers focus on collaborative cloud and edge computing for latency minimization. The research about how to evaluate the efficiency of the edge computing network is still not enough. If the efficiency of edge computing network can be evaluated, the influence factors for efficiency of edge computing can be find out to improve the design of edge computing. In order to solve the problem, we propose a Discrete-time Markov chain (DTMC) model to evaluate the efficiency of edge network by calculating the latency of the whole network. According to the experiment results, optimization strategies are given to enhance the efficiency of edge network.

The contribution of this paper is shown below:

- DTMC models are built to evaluate the efficiency of blockchain-based edge computing network.
- The influence factors (eg.the latency from edge devices to cloud) are evaluated in turn, the influence factor influence the efficiency most will be find out.
- According to the experiment results, the optimization strategies are given to enhance the efficiency of blockchain-based edge computing network.

The structure of this paper is shown as follows: section 2 gives the structure of blockchain-based edge computing, section 3 shows the basic knowledge of DCMC and the modeling process of blockchin-based cloud computing network. In section 4, the experiment results are illustrated and analyzed. Finally, the conclusion and further research are given.

II. THE STRUCTURE OF BLOCKCHAIN-BASED EDGE NETWORK

A. The framework of SDPVES

The typical architecture of blockchain-based edge network usually has 3 layers: end device layer, edge computing layer and cloud layer. The functions of each layer is shown below:

End device layer: sensors, mobile phone or PC are included in this layer, they can send or receive data from edge servers.

Edge computing layer: edge servers are included in this layer, the edge servers are also blockchin nodes in blockchin network. The edge servers can deal with some simple task and reduce the work load of cloud servers, and reflect data to the end device layer. For some complicated task, the edge servers can reflect them to the cloud layer and wait for the reflect of cloud layer. The edge servers

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00049

also help to build a blockchin network to enhance the security and information transparency of the whole network.

Cloud layer: deal with some complicated tasks and reflect the results to the edge computing layer.

The typical blockchin based edge computing network is shown in Figure 1.

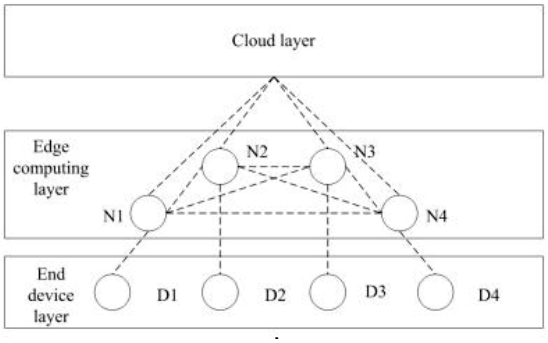


Figure 1. Architecture of blockchain-based edge computing network

In Figure 1, the meaning of each node is shown below:

- N1, N2,N3 and N4 are 4 edge servers, the edge servers can set up a blockchain network, so they can communicate with each other. They connect the cloud layer and end device layer.
- D1, D2,D3 and D4 are 4 end devices, the end devices can communicate with edge servers which is located near them.

For edge computing layer, a part of task should be connected to the blockchain network, so the consensus time should be considered. Another part of task can be deal with cloud server or edge server directly, so the consensus time is not needed. The process will be detailed illustrated in section 3.

III. BASIC KNOWLEDGE ABOUT DTMC AND THE MODELING PROCESS

A. DTMC

DTMC is a State-transition systems augmented with probabilities, DTMC has the following characters:

- States:set of states representing possible configurations of the system being modelled.
- Transitions: the transitions between states model evolution of system's state, and they occur in discrete time-steps .
- Probabilities:probabilities of making transitions between states are given by discrete probability distributions

Formally, a DTMC is a tuple (S, s_{init}, P, L) where:

- S is a set of states ("state space")
- $s_{init} \in S$ is the initial state
- $P : S \times S \rightarrow [0,1]$ is the transition probability matrix
- – L : is function labelling states with atomic propositions

According to the blockchain-based edge computing network, the state transition diagram for the DTMC model is shown in Figure 2.

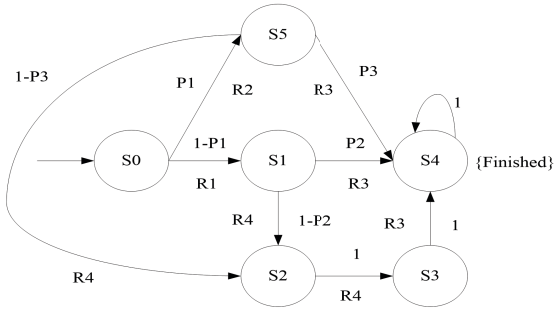


Figure 2. State transition diagram for the DTMC model

The parameters for Figure 2 are explained below:

- S0 is the initial state, end devices prepare to send messages to edge computing servers.
- S1 is the state that the messages come to edge computing servers, and the consensus time in blockchin network is not needed (means the messages are not necessary to be added to blockchin network)
- S2 is the state that the messages come to the cloud servers.
- S3 is the state that messages come to edge computing servers from cloud.
- S4 is the state that messages come to end device layer from edge computing layer.
- S5 is the state that the messages come to edge computing layer, but the consensus time in blockchain network is needed
- P1 is the probability that the consensus time in edge computing layer is needed. 1-P1 means the probability the consensus time is not needed.
- P2 is the probability that the messages can be dealt by edge computing servers.1-P2 means the probability messages should be dealt by cloud servers.
- P3 is the probability that the messages can be dealt by edge computing servers after consensus.1-P3 means the probability messages should be dealt by cloud servers .
- R1 stands for the latency between S0 and S1.
- R2 is the latency between S0 and S5.
- R3 is the latency between S1/S3/S5 and S4.
- R4 is the latency between S1/S5 and S2 or between S2 and S3.

B. Modelling Process

In this paper, DTMC is used to evaluate the efficiency

of blockchain-based edge computing, the modeling process is show below.

Firstly, we can give values to P1, P2 and P3, in experiments, the value can be changed.

```
const double P1=0.1;
const double P2=0.2;
const double P3=0.2;
```

Secondly, the state transition can be modeled.

```
module edge
s : [0..6] init 0;
[] s=0 -> 1-P1: (s'=1) + P1: (s'=5);
[] s=1-> 1-P2 :(s'=2)+P2 :(s'=4);
[] s=4 -> (s'=4);
[] s=2 -> (s'=3);
[] s=3 -> (s'=4);
[] s=5 -> 1-P3 :(s'=2)+P3 :(s'=4);
endmodule
```

Finally, the latency between different states can be set in software PRISM by adjusting the rewards.

When the transition system reaches the last state S4, the whole process finishes. The average latency can be calculated.

We can set P1=P2=P3=0.1, R1=R3=2s,R2=4s, R4=3s, then we can use theory calculation method to verify the correctness of our DTMC model. The experiment result of DTMC model is shown in Figure 3. The result shows that the average latency is 9.6s.

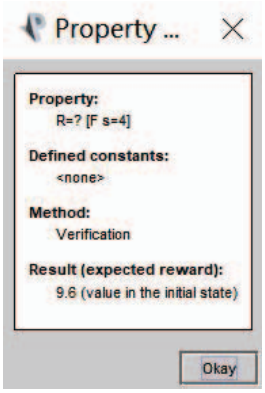


Figure 3. Average latency (by using DTMC model)

Then we use theory calculation method to calculate the average latency L:

$$L = R1(1-P1)+(R4+R4+R3)(1-P1)(1-P2)+R3(1-P1)P2+R$$
$$(R4+R4+R3)P1(1-P3)+R3P1P3 = 2\times0.9+8\times0.9\times0.9+2\times0$$
$$+4\times0.1+8\times0.1\times0.9+2\times0.1\times0.1=9.6$$

The result is same with previous one, so our DTMC model is verified to be correct.

IV. Experiment Results

Before experiment, the parameters should be determined For each experiment. The parameters are shown in Table 1. In this paper, duo to the time limitation. Influence of P3 and R1 is not included in the experiments.

TABLE I. Parameters

	P1	P2	P3	R1	R2	R3	R4
Experiment 1	0.05,0.1,0.15,0.2	0.2	0.2	2	4	2	3
Experiment 2	0.1	0.1,0.15,0.2,0.25,0.3	0.2	2	4	2	3
Experiment 3	0.1	0.2	0.2	2	3,4,5,6,7	2	3
Experiment 4	0.1	0.2	0.2	2	4	1,2,3,4,5	3
Experiment 5	0.1	0.2	0.2	2	4	2	2,3,4,5,6

A. *Influence of P1*

For experiment 1, the experiment result is shown in Figure 4. Figure 4 reflects the influence of P1 about the efficiency of the blockchain-based edge computing network.

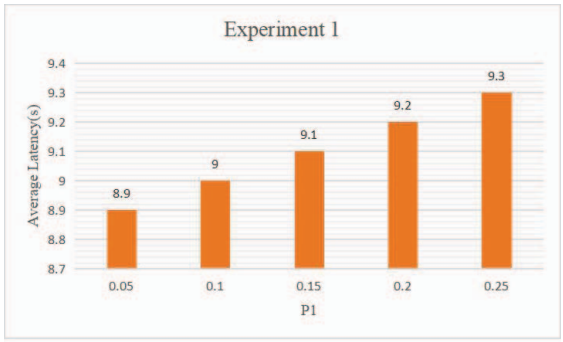


Figure 4. Experiment 1

Remark. Average latency is proportional to P1. However, the influence is not very obvious.

B. *Influence of P2*

For experiment 2, the experiment result is shown in Figure 5. Figure 5 reflects the influence of P2 about the efficiency of the blockchain-based edge computing network.

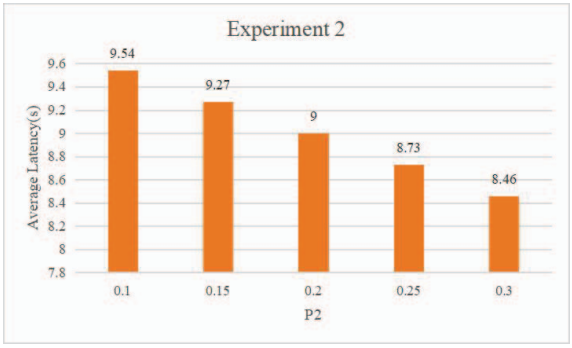


Figure 5. Experiment 2

Remark. Average latency is inversely proportional to P2. The impact of it is lager than P1.

C. *Influence of R2*

For experiment 3, the experiment result is shown in Figure 6. Figure 6 reflects the influence of R3 about the efficiency of the blockchain-based edge computing network.

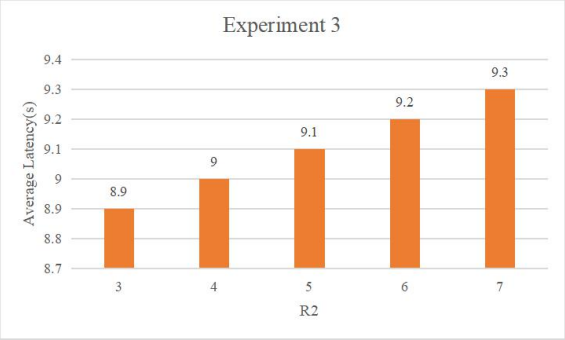


Figure 6. Experiment 3

Remark. Average latency is proportional to P1. However, the influence is not very obvious.

D. *Influence of R3*

For experiment 4, the experiment result is shown in Figure 7. Figure 7 reflects the influence of R3 about the efficiency of the blockchain-based edge computing network.

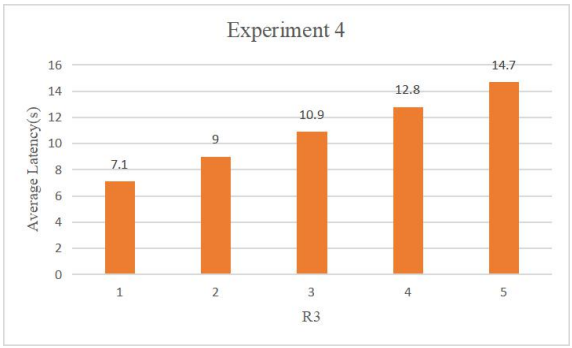


Figure 7. Experiment 4

Remark. Average latency is also proportional to R3. The impact of it is lager than R2.

E. *Influence of R4*

For experiment 5, the experiment result is shown in Figure 8. Figure 8 reflects the influence of R4 about the efficiency of the blockchain-based edge computing network.

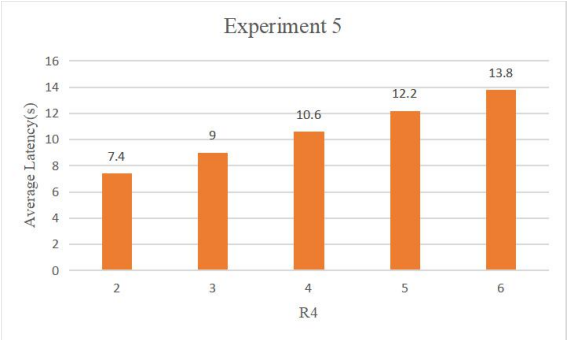


Figure 8. Experiment 5

Remark. Average latency is also proportional to R4. The impact of it is smaller than R2. For R2,R3 and R4, the impact of R3 is largest.

V. CONCLUSION

In this paper, a DTMC model is supposed to evaluate the efficiency of blockchain-based edge computing network. The experiment results are given for analyzing. For P1 and P2, the impact of P2 is larger. For R2,R3 and R4, the impact of R3 is largest. If we want to improve the efficiency of blockchain-based edge computing network, the service ability of edge computing server should be improved to reduce the probability the messages should be dealt by cloud. What's more, the latency between edge computing layer and end device layer should be reduced.

However, the modeling is not very complicated. In further research, I want to explore more complicated models to meet the evaluation need of real case. These will be included in further research.

ACKNOWLEDGMENT

The paper is supported by Natural Science Foundation of Jiangsu Province No. 18KJB413009

REFERENCES

[1] A, Osama Sohaib , et al. "Cloud computing model selection for e-commerce enterprises using a new 2-tuple fuzzy linguistic decision-making method." Computers & Industrial Engineering132(2019):47-58.

[2] Zhao-Peng C . Preliminary Study on Power System Cloud Computing[J]. Digital Technology & Application, 2019.

[3] Munoz R , Vilalta R , Yoshikane N , et al. Integration of IoT, Transport SDN, and Edge/Cloud Computing for Dynamic Distribution of IoT Analytics and Efficient Use of Network Resources[J]. Journal of Lightwave Technology, 2018:1420-1428.

[4] Turkanovic M , Holbl M , Kosic K , et al. EduCTX: A blockchain-based higher education credit platform[J]. IEEE Access, 2018:5112-5127.

[5] Alshamsi A , Andras P . User Perception of Bitcoin Usability and Security across Novice Users[J]. International Journal of Human-Computer Studies, 2019, 126.

[6] [6]Tatiana, G, Choleva, et al. EdgeChain: An Edge-IoT Framework and Prototype Based on Blockchain and Smart Contracts[J]. Talanta, 2019.

[7] Ardakani A , Condo C , Gross W J . Fast and Efficient Convolutional Accelerator for Edge Computing[J]. IEEE Transactions on Computers, 2020, 69(1):138-152.

[8] Wei H , Feng W , Zhang C , et al. Creating Efficient Blockchains for the Internet of Things by Coordinated Satellite-Terrestrial Networks[J]. IEEE Wireless Communications, 2020, 27(3):104-11.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Development of Classroom Assistant Software Based on Mini Program

Yucheng Guo
School of Computer Science
Wuhan University of Technology
Wuhan, China
guoyucheng@whut.edu.cn

Yufei Liang
Yangtze Optical Fiber and Cable Joint Stock Limited Company
Wuhan, China
119071512@qq.com

Abstract— **This paper introduces a classroom assistant application based on mini program which provides classroom teaching interaction, attendance checking, mobile phone management and other functions. Through this application, teachers can interact with students in class and ask them questions to know how well they have mastered the knowledge; they can initiate attendance check to understand students' attendance in an efficient and time-saving way; they can release assignments and notifications to students in the form of push messages; the application has developed a software activity management function. This function can monitor whether students play with their cell phones in class and thus can avoid student distractions improving the quality of teaching.**

Classroom assistant; mini program; WeChat; phone status monitoring;

I. INTRODUCTION

Currently, many university courses are taught in large groups, with more than 100 students in 3 or 4 classes at a time. In this case, how to manage the class effectively becomes a headache. First of all, it is too costly to use valuable classroom teaching time if roll call is done frequently, and it is not conducive to managing students' attendance if it is not done; secondly, due to the large number of students, classroom teaching generally lacks interaction, and there are often appearing the situations which the students are keeping silence when the teacher asks questions; at the same time, due to the popularity of cell phones, tablets and other smart devices, students often use these devices to read novels, chat with friend or even play games in class, which seriously affects the teaching effectiveness. At the same time, due to the popularity of smart devices such as mobile phones and tablets, students often use these devices to read novels, chat or even play games in class, which seriously affects the teaching effect. In an era when smart devices are so popular, it is worth exploring how to not let these devices affect the quality of teaching, but to use them as an effective aid to classroom teaching [1].

II. MINI PROGRAM OF THE WECHAT

This paper introduces a classroom assistant application based on mini program which uses smart phones as teaching aids thus avoiding the problems mentioned above. Mini program is a functional component of WeChat. WeChat is one of the most used mobile apps in China, with a huge market of over 350 million daily active users and over 1.2 billion monthly active users. In 2017, WeChat officially launched mini program, allowing external developers to run their own code and conduct business inside WeChat. This sparked an overwhelming response, and as of June 2021, the number of mini programs has exceeded 6.5 million. A mini program is an application inside WeChat, where external code runs inside WeChat, a mobile app, through a form called mini program. However, it is more accurate to say that mini program can be considered a website that can only be opened and browsed with WeChat. The technical model for mini program and web pages is the same, and the JavaScript language and CSS styles used are the same, except that the HTML tags for web pages have been slightly modified to WXML tags. Mini program pages are essentially web pages. The special thing about mini program is that although they are web pages, they do not support browsers, and all browser APIs are not available, only those provided by WeChat [2].

The biggest advantage of the mini program is that it is based on WeChat. And WeChat is cross-platform, so mini program can run under different hardware environments. Most of the functions of WeChat App, such as taking photos, scanning, payment, etc., can be used by mini program. WeChat provides a variety of packaged APIs for developers to use. Developers don't have to implement them themselves, and they don't have to consider the platform differences between iOS and Android. Therefore, in this paper, we choose to use mini program for development [3].

III. THE KEY CONTENT OF THE APPLICATION DESIGN

This paper introduces a mini program-based classroom assistant application. The application provides functions such as classroom teaching interaction, quick attendance checking, and classroom cell phone management. First of all, the application allows teachers to post questions in class through a simple method, so that students can quickly answer with their smartphones, avoiding the situation that the teacher asks questions and students are all keeping silent; at the same time, students can also provide timely feedback to the teacher through this system on whether the lecture progress is too fast, whether a certain point of knowledge is understood, etc., so that teachers and students can interact well and effectively in class, thus can improve the quality of teaching. Secondly, teachers can call the roll easily and quickly by using this application. The roll call is initiated by the teacher and completed by students in a limited time using their cell phones or other smart devices. By using AGPS technology, cell phone geolocation information is added to verify whether students are within the specified range when the roll calling, thus ensuring the accuracy of attendance checking. Finally, the application has developed a software activity

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00050

monitor function. This function asks students to open this classroom assistant application in class to assist in classroom teaching; it sends alerts to the server when students switch this software to the background and displays it on the teacher's client, which allows teachers to know whether students are paying attention to class and facilitates teaching management. The classroom assistant application makes full use of the latest achievements in the development of modern technology, digs deeper and gives full play to the functions of smart devices such as smart phones, so that student's the smart devices are no longer the distracting toys, but the powerful tools to aid learning and work. Using the system well will further enhance the level of teaching information management and is of great significance to effectively improve the quality of teaching services.

The main contents of The Classroom Assistant application development include: database server, backend management system, student version mini program, teacher version mini program. Same as other web applications, the general idea is that the user sends a response operation request, the server receives the request, then performs the relevant operation on the database, and finally feeds back the response operation result to mini program clients. But all of this are doing in mini program of WeChat. So, here is a little different from the previous process. In this paper, we introduce the three functions of the classroom assistant application: silent registration, roll call with locating, and cell phone status monitoring.

A. Silent Registration

There are several common ways to register and login to general software.

- Customized account and password registration
- Email address registration
- Mobile phone number registration

Since this software is based on WeChat mini program, we use WeChat account login and access token method. By using this method, we can reduce the registration and login process, and finish the registration and login in a more senseless way to achieve a "silent" effect. For further access to user nicknames, cell phone numbers and other information, for user privacy reasons, users need to actively agree to authorize. Only then can this information be obtained. In WeChat, due to the different sensitivity of different users, the system provides "authorization" in different ways for different user information.

- Call the specific API way, pop-up authorization: if the user is not authorized, the address authorization interface will pop up; if rejected, it will not pop up again, directly return to failure.
- Explicit buttons: just like <button open-type="xxx" />. Only Suitable for user sensitive information, user cell phone number etc., need to cooperate with back-end symmetric encryption and decryption to get the data. If the user has rejected, click the button next time, it will still pop-up window.
- via wx.authorize() API: Ask for authorization in advance, and then don't need to pop up again when you need to get relevant information. So, you can get something in 'silence' when user get in your mini program.

The classroom assistant application adopts the third authorization mode in registration and login, and the first mode was adopted in the roll call part. The specific implementation flow is as figure 1.

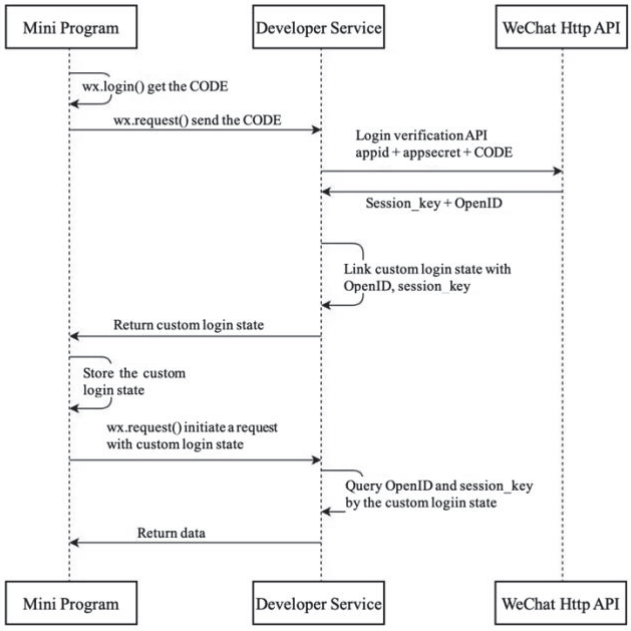


Figure 1. The registration process in mini program

The steps are divided into three parts. First, the front-end obtains the one-time encrypted credential CODE via wx.login() and gives it to the back-end. Then, the back-end transfers the CODE to the WeChat server in exchange for the unique user identifier OpenId and the authorization credential session_key. Session_key is used for subsequent special API calls between the server and the WeChat server. Finally, the backend transmits the user credentials obtained from the WeChat server to the front-end with the access token generated by the backend. The front-end saves it and brings it to the back-end when next requested to identify the specific user.

B. roll-call with location

In order to ensure the accuracy of roll call, it is required to open GPS positioning. In order to successfully and accurately acquire geographic locations, The mini program need require three steps of authorization verification.

- System level, system switch for geolocation , whether system GPS switch is on
- WeChat App side, switch to allow WeChat to use location , whether WeChat has permission to obtain system location
- Mini program side, whether the mini program is authorized by the user to use geolocation , whether the mini program side requests permission from WeChat

Specifically in the coding, we need modify app.json of mini program project fist, add code at the same level as pages, and use the wx.getLocation method just like below.

```
onShow: function() {
  let sInfo = wx.getSystemInfoSync();
  this.setData({
   locationEnabled: sInfo.locationEnabled + "",
```

```
      locationAuthorized: sInfo.locationAuthorized + ""
    })
    let that = this;
    wx.getLocation({
      type: 'wgs84',
      success(res) {
        that.setData({
          locationMini: "true",
          location: JSON.stringify(res)})
      },
      fail(err) {
        that.setData({
          locationMini: "false",
          location: "Mini program is not authorized."
        })
      }
    })
  },
```

call wx.getSystemInfoSync() to determine whether system positioning is open and whether WeChat is authorized, if both have been opened and then determine whether the mini program has obtained the location permission, and then you can get a more accurate location. After getting the precise location, compare the student check-in location with the teacher's location. If it is within the specified range, in this application the range is set to 300 meters according to the accuracy of WeChat positioning, then the roll call is successful, otherwise it is not successful.

C. smart phone status monitoring

In order to avoid students playing with their smart phones during classes, it is necessary to manage the behavior of students' smart phones. Cell phone status monitor is a difficult problem, and many developers have used accelerometer and gyroscope listening methods to solve it. However, this method is slightly complicated and prone to misjudgment. For this reason, instead of using accelerometer and gyroscope, this paper adopts the method of listening to whether the user leaves this program or not to determine whether the student's cell phone leaves this program to do other things. In mini program there is an onHide() function which provides similar functionality [4,5,6].

This function is easy to use but the disadvantage is also very obvious. According to WeChat official API manual, The condition for the onHide() function to be triggered is "when the mini program enters into the background from the foreground ", that is, as long as the user cannot see the application, including leaving the mini program, leaving this page, leaving WeChat, extinguishing the screen and etc., the onHide() will be called. So how do you tell if onHide is called when you leave the page, or if it just turns off the screen. This is a very critical question. Through research we found that that the mini program also provides another function to get the screen brightness, that is wx.getScreenBrightness(), as its name this function can get the brightness value of the device screen. Therefore, we only need to call the getScreenBrightness() function in the onHide function to determine whether the students' cell phones has left the application, thereby achieving the capture of cell phone cut away from the application events and give hints. By use this method the smart phone status monitoring is achieved which solves the problem of students playing with cell phones in class in certain extent and makes classroom management easier.Pseudo code for the smart phone status monitoring as below.

```
  onHide: function() {
    wx.getScreenBrightness({
      success: function(res) {
        if (res.value !== 0) {
          $wxDialog().alert({
            resetOnClose: true,
            maskClosable: false,
            closable: false,
            title: 'Error',
            content: "This task failed because you
exited the page during the task.",
            onConfirm(e) {
              wx.switchTab({url: "../today/results"})
            },
          })
        }
      }
    });
  }
```

IV. CONCLUSION

The classroom assistant application introduced in this paper is developed based on mini program which combining mobile computing technology with teaching, solving the current teaching purely on the large classroom class roll call takes too much time, classroom teaching lacks interaction, and students' cell phones are difficult to manage in class. Taking full advantage of WeChat's market share and the ease of use of small programs, the classroom assistant application can be easily used cross-platform, instead of downloading the application, users can simply tap on the applet link and use it. So it is easy to use and very attractive.

REFERENCES

[1] Rodriguez G, Soria Á, Campo M. Virtual Scrum: A teaching aid to introduce undergraduate software engineering students to scrum[J]. Computer Applications in Engineering Education, 2015, 23(1): 147-156.

[2] Cheng K, Schreieck M, Wiesche M, et al. Emergence of a Post-App Era-An Exploratory Case Study of the WeChat Mini-Program Ecosystem[C]//Wirtschaftsinformatik (Zentrale Tracks). 2020: 1444-1458.

[3] Hao L, Wan F, Ma N, et al. Analysis of the development of WeChat mini program[C]//Journal of Physics: Conference Series. IOP Publishing, 2018, 1087(6): 062040.

[4] Yurur O, Liu C H, Moreno W. Unsupervised posture detection by smartphone accelerometer[J]. Electronics Letters, 2013, 49(8): 562-564.

[5] Tlili F, Haddad R, Bouallegue R, et al. A Real-time Posture Monitoring System Towards Bad Posture Detection[J]. Wireless Personal Communications, 2021: 1-21.

[6] Vo Q V, Lee G, Choi D. Fall detection based on movement and smart phone technology[C]//2012 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future. IEEE, 2012: 1-4.

Research on Amazon Forest Fire Based on Cellular Automata Simulation

Weizheng Sun,Wenxiang Wei*
Hunan University of Science and Technology
School of Information and Electrical Engineering
Hunan, China
981929626@qq.com,Ytwei010702@126.com

Jingyu Chen,Kaiqi Ren
Hunan University of Science and Technology
School of Information and Electrical Engineering
Hunan, China
chen542196811@icloud.com,bf2bf3@126.com

***Abstract*—Wildfire, especially uncontrolled flammable vegetation hazard in rural or wilderness areas, seriously affected human life and safety, prediction methods has been researched widely in recent decades. According to Amazon forest fire situation, an advanced system hierarchical clustering model is established in this paper, and the diffusion simulations of Amazon forest fire is implemented by using the cellular automata model, which constructed with the previous model. Then, the diffusion speed and diffusion range of different brightness parameters are analyzed. Furthermore, combined with the three-dimensional effect of fire and smoke simulated by particle system, the entropy weight TOPSIS model is constructed to the optimized ranking of wildfire risk index with its longitude and latitude, and obtain the relationship diagram of latitude, longitude and wildfire risk index after extracting data. The feasibility and rationality of the proposed method is verified by simulation results.**

Keywords-advanced system hierarchical clustering model; cellular automata; entropy weight TOPSIS; outlier analysis; particle system simulation

I. INTRODUCTION

Extreme events occur for wildfire from time to time, although the burning area of only 3-5% of wildfires exceeds 100 hectares[1]. However, the burning area of the top 1% of wildfires accounts for 80-96%. Nowadays, communities around the world are currently at risk of massive, severe wildfires that directly threaten people and bring life-threatening impacts on air quality, water availability and quality, as well as soil integrity. After the devastating wildfire season in the United States in 2000, a new term "megafire" was coined. The impact or severity of wildfires on society has reached a new level. In fact, in Western states, the frequency of wildfires has been increasing. Statistical analysis of a wide range of areas shows the relationship between the acres burned and its intuitively favorable conditions for fire occurrence. These views lead to the general attribution of large-scale fires[2]. More speculatively, individual fires are attributed to climate change and fuel accumulation. The understanding and prediction of large-scale wildfire events in the world is a developing interdisciplinary research field, which develops rapidly with the development and application of computational models. Recent two-way coupled computational fluid dynamics models, including weather forecasting models and modules, are equipped with algorithms representing fire spread and heat release, simulating the interactions of the fire atmosphere across three orders of magnitude.(The data used in this article comes from this website:http://mcm.tzmcm.cn/).

At present, fire diffusion prediction is used mostly in indoor, ships, high-rise buildings and other scenes, but rarely in natural scenes. The main causes lie in that many factors in the natural environment lead to a complex and nonlinear change process of fire diffusion. Yang Jun et al simulated land-use change through the cellular automata model, and the experimental results were accurate[3]. Therefore, cellular automata are suitable for complex system simulation. According to these, this paper proposes a method to study forest fire diffusion in Amazon based on cellular automata simulation. This method avoids the subjectivity of data, that is, no objective functions and tests are needed. Also, it can well describe the comprehensive influence intensity of multiple indexes

II. ESTABLISH THE HIERARCHICAL CLUSTERING MODEL OF ADVANCED SYSTEM

According to the elbow method, we can roughly estimate the optimal number of clusters through graphs.

Supposing that n samples in total were divided into *K* classes ($k \le n-1$, that is, there is at least a class with two elements), C_k was used to define the k-th class (*k=1,...,K*), and u_k was used to refer to its center of gravity. The degree of distortion is known for each class of the class is equal to the sum of squares of the distance from the center of gravity and its position of internal members, then the distortion of the k-th class is as follows:

$$\sum_{i \in C_k} |x_i - u_k|^2 \tag{1}$$

The total distortion degree (polymerization coefficient) of all classes was defined as:

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} |x_i - u_k|^2 \tag{2}$$

Supposing the horizontal coordinate as the number of clustering classes K, the vertical coordinate as the polymerization coefficient J, then the polymerization coefficient line graph was made and shown in Fig. 1:

Identify applicable sponsor/s here. *(sponsors)*

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00051

As shown in Fig. 1, a significant change occurs at K=6, and the number of clustering classes was determined as 3.

After determining the number of clustering classes, the new class sequentially combined by two classes with short distance constantly iterates, and the final classification result was shown in the ice chart:

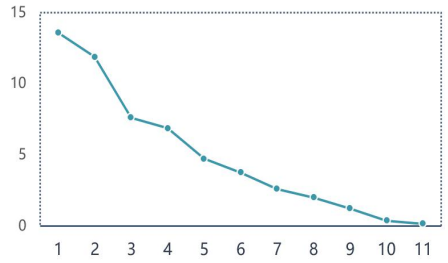


Figure 1. Polymerization coefficient line graph

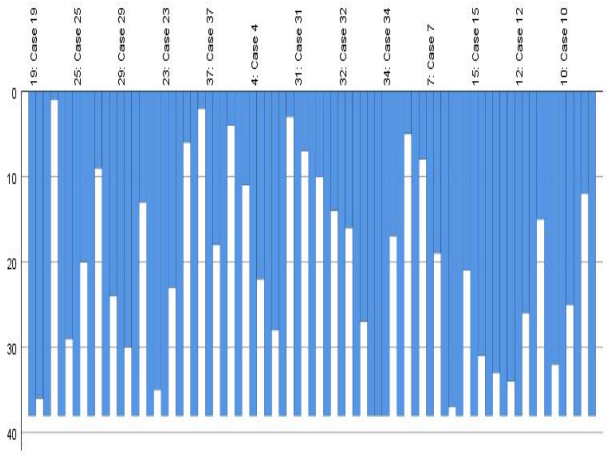


Figure 2. Ice chart

Therefore, the data extraction section can be substantially clustered into 4 groups, that is, four sets of variables. The four sets of variables were substituted into the correlation coefficient equation:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{3}$$

Since there may be an over-fitting phenomenon in the calculation process, so the cross-test was conducted.

TABLE I. CROSS-TEST RESULTS TABLE

CV Accuracy Scores	0.987	0.965	0.943	0.935	0.998	0.924	0.976	0.934	0.937	0.978
CV Accuracy	0.95	+/-	0.035							

As shown in Table 1, the cross-test accuracy of this model fluctuates between 0.950.035, indicating that the model test results are good.

III. FIRE POINTS DISTRIBUTION

A confidence interval of more than 80% was chosen and points planning with distance less than 1km within 24 hours were proposed as follows.

$$\begin{cases} T_{20} \geq 315 \\ 20 + T_S \leq T_{20} \leq 70 + T_S \\ T_{20} - T_{31} \geq 18 \end{cases} \tag{4}$$

T_{20} and T_{31} were used as the brightness of channels 20 and 31; TS was used for the ground temperature with the unit of K.

IV. PARTICLE SYSTEM SIMULATION

The three-dimensional effect of fire and smoke was simulated using a particle system.As a professional three-dimensional simulation platform, Unity3D itself provides a rich library of vegetation and particle systems, which provided the conditions for the simulation of flame and smoke. Steps of realizing the visual three-dimensional effect of fire and smoke using Unity 3D were shown as follows:

A. Establish templates of fire and smoke particles with a given particle life cycle, size, color and transparent Alpha channel value.

B. Set up the location of the fire and smoke particles and initialize the emission rate and the number of particles through simulating the regional digital terrain information.

C. Control the generation and die of particles using life cycle according to the forest fire spread speed and accordingly update the apparent characteristics and attributes to realize the visual three-dimensional dynamic movement of fire and smoke.

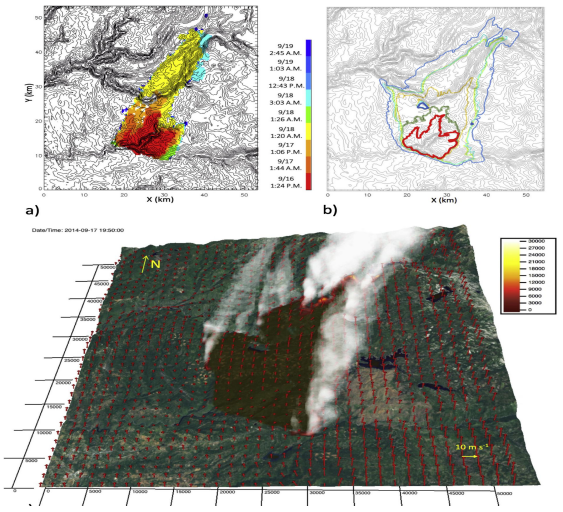


Figure 3. Visual three-dimensional dynamics of fir and smoke based on particle system.

From Fig. 5, it can be seen that the relationship between these two parameters and wildfire index is not only valid, but positive and negative, in which two indexes are used to construct the cell space.

V. CELLULAR AUTOMATA & ENTROPY TOPSIS MODEL

MATLAB was used to demonstrate Amazon forest fires and analyze them spatially.The cellular automata model of forest fire has three states: vacancy, burning trees and trees. Then the state of a cell at the next moment is determined by its state at that moment and the states of four surrounding neighbors according to certain rules, and the rules are set as follows:

- If four neighbors of a tree cellular are burning, then the state of the cellular at the next moment is still burning.
- A burning cellular becomes the number of vacancies (n) at the next moment.
- All tree cellular start burning with a low probability (to simulate the fire P caused by lightning).
- All vacant cellular become trees with a low probability (to simulate the growth of new trees).
- Fire brightness (T) detected by Brightness Satellite

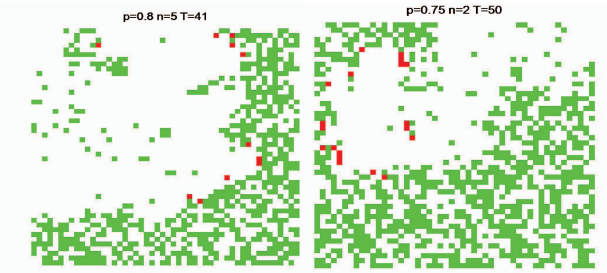


Figure 4. Distribution diagram of the wildfire risk index

The higher the probability of lightning splitting trees is and the more the surrounding vacancies are, the more intense the fire hazard will be and the higher the wildfire risk index will be.

The advanced simulation effect diagram after adding wind speed and brightness is shown as follows:

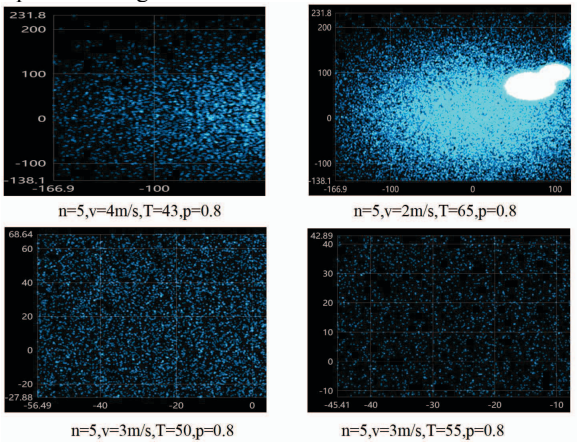


Figure 5. Advanced simulation effect diagram after adding wind speed and brightness

Through cleaning the six sets of simulation charts, the spatial changes of fire diffusion. The stronger the brightness and wind detected by satellite are, the higher the wildfire risk index and the faster the spatial diffusion of fire will be; The more intensive the fire points are, the weaker the brightness and the spatial diffusion, and the lower the wildfire risk index will be.

Then, the entropy weight TOPSIS model was constructed to sort the wildfire risk index in different latitude and longitude spaces according to the simulation results. After standardizing the data, the TOPSIS method was used to construct positive and negative ideal solutions; and then according to multiple attributes, comprehensive evaluation values are given to regions to divide regional weights.

A. Construct the scheme set, attribute set and decision matrix[6].

Supposing the scheme set as $S=\{S_1,S_2,...,S_m\}$, the attribute set as $P=\{P_1,P_2,...,P_n\}$, the attribute set including benefit attribute, that is, the bigger the attribute value is, the better the index will be, and cost attribute, that is, the smaller the attribute value is, the better the index will be; as well as the multi-attribute decision-making matrix $X=(x_{ij})_{m\times n}$, of which x_{ij} is the index value of the *j*-th attribute of the *i* scheme, $i=1,2,...m, j=1,2,...,n.$

B. Solve the standardization decision-making matrix.

The method of vector planning was used to design the standardization decision-making matrix $Y=(y_{ij})_{m\times n}$, among them:

$$y_{ij}=\frac{x_{ij}}{\sqrt{\sum_{i=1}^{m}x_{ij}^2}}, i=1,2,...,m, j=1,2,...,n \tag{7}$$

C. Determine the factor weight.

The characteristic proportion of the index is:

$$f_{ij}=\frac{y_{ij}}{\sum_{j=1}^{n}y_{ij}} \tag{8}$$

The information entropy is:

$$H_i=-\frac{1}{\ln n}\sum_{j=1}^{n}f_{ij}\ln f_{ij} \tag{9}$$

Among them *ln0=0.*

The entropy weight method can be effectively used to cover the evaluation of variation degree and objectively reflect its importance. The entropy is calculated as follows:

$$w_i=\frac{1-H_i}{m-\sum_{i=1}^{m}H_i} \tag{10}$$

The weight is calculated as:

TABLE II. INDEX WEIGHT

latitude	longitude	brightness	confidence	bright_t31
0.21	0.123	0.121	0.1181	0.1748

After constructing weighting matrices $Z=(z_{ij})_{m\times n}$, the weight vector of various factors was given by the entropy weight method $w=[w_1,w_2,...,w_n]^T$, then

$$z_{ij}=y_{ij}\cdot w_i, i=1,2,...,m, j=1,2,...,n \tag{11}$$

D. Determine the positive ideal solutionand negative ideal solution.

Supposing $A^+=(z_1^+,z_2^+,...,z_n^+), A^-=(z_1^-,z_2^-,...,z_n^-)$, among them

$$z_j^+=\begin{cases}\max\limits_{1\le i\le m} z_{ij}, j\in j^+\\ \min\limits_{1\le i\le m} z_{ij}, j\in j^-\end{cases}, z_j^-=\begin{cases}\min\limits_{1\le i\le m} z_{ij}, j\in j^+\\ \max\limits_{1\le i\le m} z_{ij}, j\in j^-\end{cases} \tag{12}$$

Results and scores are shown as follows:

E. Calculate the distance from various schemes to ideal solutions.

Distance between the scheme S_i to the positive ideal solution A^+.

$$T_i^+, T_i^+=\sqrt{\sum_{j=1}^n (z_{ij}-z_j^+)^2}, i=1,2,...,m \tag{13}$$

Distance between the scheme S_i to the positive ideal solution A^-.

$$T_i^-, T_i^-=\sqrt{\sum_{j=1}^n (z_{ij}-z_j^-)^2}, i=1,2,...,m \tag{14}$$

F. Calculate all ranking index values of each scheme.

The comprehensive evaluation index is:

$$C_i=\frac{T_i^-}{T_i^+ + T_i^-}, i=1,2,...,m \tag{15}$$

Results and scores are as follows:

latitude	longitud	the risk ofwild fire
-12.311	-48.151	0.003386365
-12.073	-48.735	0.001993221
-12.14	-55.078	0.002064897
-10.723	-49.527	0.002372153
-11.264	-53.951	0.004263493
-6.705	-45.571	0.006053936
-6.323	-53.409	0.006065392
-6.325	-53.421	0.003587771
-6.707	-45.585	0.006054172
-6.116	-52.585	0.005147443

Figure 6. Quantitative map of longitude and latitude

All extracted visual data of longitude, latitude and risk were shown in the following relationship diagram of latitude, longitude and wildfire risk index of the Amazon forest fires:

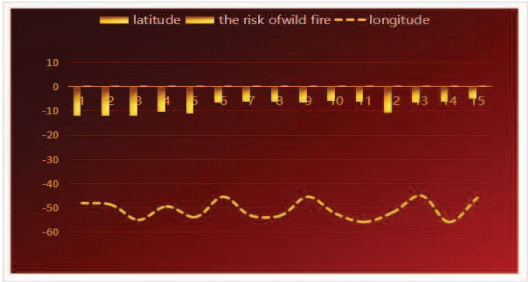


Figure 7. Relationship diagram of latitude, longitude and wildfire risk index

VI. CONCLUSIONS

Based on the advanced system hierarchical clustering model, this paper constructs the cellular automata model, and obtains the diffusion simulation results of the Amazon forest fire. Also, the diffusion speed and range of different brightness parameters are analyzed. Then, combined with the three-dimensional effect of fire and smoke simulated by particle system, the entropy weight TOPSIS model is constructed to the optimized ranking of wildfire risk index and longitude and latitude and obtain the relationship diagram of latitude, longitude and wildfire risk index after extracting data.

ACKNOWLEDGMENT

A Project Supported by Scientific Research Fund of Hunan Provincial Education Department(18A195);Post-competition research on question B of the 9th "Certification Cup" Mathematics Chinese Mathematical Modeling International Competition in 2020.

REFERENCES

[1] *D.E. Calkin, K.M. Gebert, J.G. Jones, R.P. Neilson, Forest Service large fifire area burned and suppression expenditure trends, 1970−2002, J. For. 103 (2005) 179−183.*

[2] *D. Strauss, L. Bednar, R. Mees, Do one percent of forest fifires cause ninety-nine percent of the damage? For. Sci. 35 (1989) 319−328.*

[3] *K. Short, A spatial database of wildfifires in the United States, 1992−2011, Earth Syst. Sci. Data Disc. 6 (2013) 297−366.*

[4] *J. Williams, Exploring the onset of high-impact mega-fifires through a forest land management prism, For. Ecol. Manage. 294 (2013) 4−10.*

[5] *A.L. Westerling, H.G. Hidalgo, D.R. Cayan, T.W. Swetnam, Warming and earlier spring increase western US forest wildfifire activity, Science 313 (2006) 940−943.*

[6] *Le Z, Li XIN. A method to analyze the fire spread of historical settlements and their value loss based on the simulation of metacellular automata: an example of the historical and cultural village of China, Nanjing Qixiao Village[J]. Architecture and Culture,2021(01):114-116。*

[7] *Sari Fatih. Forest fire susceptibility mapping via multi-criteria decision analysis techniques for Mugla, Turkey: A comparative analysis of VIKOR and TOPSIS[J]. Forest Ecology and Management,2021,480.*

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Improved BiLSTM Model For Online Food Safety Risk Prediction

1st Zekai Zhang
School Of Artificial Intelligence and Computer Science,Jiangnan University
WuXi ,China
E-mail: 971988947@qq.com

3rd Jun Sun
School Of Artificial Intelligence and Computer Science,Jiangnan University
WuXi ,China
E-mail: junsun@jiangnan.edu.cn

2nd Mao Li
School Of Artificial Intelligence and Computer Science,Jiangnan University
WuXi ,China
E-mail: wxmaoli@163.com

3rd Qidong Chen
School Of Artificial Intelligence and Computer Science,Jiangnan University
WuXi ,China
E-mail: cqd_mu@hotmail.com

***Abstract*—The prediction of security risk is a new problem in the field of modeling. The value of food safety risk is dynamic, non-linear and non-stationary. In order to improve the prediction accuracy of non-stationary sequences, this paper proposes an improved time series prediction model of bidirectional long short-term memory network based on NS-ADAM optimization algorithm. By using NS-ADAM optimization algorithm, the non-stationary series can be approximately regarded as stationary series in a fixed window length. Combined with the advantages of two-way long-term and short-term network model to strengthen the correlation between current data and historical data, the combined network model significantly improves the prediction accuracy of non-stationary time series. The experimental results show that compared with the traditional integrated moving average autoregressive model and the conventional long-term and short-term memory network model, the improved model has higher prediction accuracy and better fitting effect for non-stationary series.**

Keywords-network food safety risk prediction;optimization algorithm;bidirectional long short-term memory network

I. INTRODUCTION

In recent years, food safety issues [1] have always maintained a high degree of hot spots at home and abroad. With the rapid development of the Internet economy, the proportion of online ordering methods in the consumer's dietary consumption patterns has shown a leap-forward growth, but the corresponding systems and laws and regulations of the online ordering industry have not been able to adapt to the rapid development of the industry. Online food safety issues continue to emerge, arousing increasing social concern. Most scholars at home and abroad have studied such issues mainly from the perspectives of social management and economics[2], and put forward their thoughts on strengthening the supervision of the online food ordering industry from the three aspects of the government, the food delivery platform[3], and the consumer. And suggestions. However, there is almost no research on the risk prediction of online food safety based on the modeling field. This article takes the long-term food safety risk value of a certain business in a city as an example, uses neural network methods to predict online food safety risks [4], and provides technical support for online food safety issues from an objective and scientific point of view.

II. BACKGROUND INTRODUCTION

A. LongShort-term Memory Network

LSTM[5-7] is extended on the basis of RNN[8], using "Gating" unit to keep the model and the long-term dependence of the learning input. The LSTM model consists of three gates, called Input Gate, Output Gate, and Forget Gate. The Forget gate is used to retain or remove the existing information, the Input Gate specifies the extent to which the new information enters the memory, and the Output Gate finally determines the value we will output based on all the information currently received. The gate operation is mainly through a Sigmoid neural layer and a point-by-point multiplication operation to achieve the screening of information.

B. Bidirectional LongShort-term Memory Network

BiLSTM[9-10] is an extended model of LSTMs, in which there are two LSTMs in the input data part. In the first round, LSTM is applied to the input sequence. In the second round, the inverse form of the input sequence is entered in the LSTM model. Applying LSTM twice can improve the long-term dependence performance of learning, thereby improving the accuracy of the model.

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00052

III. IMPROVED BILSTM BASED ON NEW OPTIMIZATION ALGORITHM

A. Adaptive Moment Estimation and Optimization Algorithm NS-ADAM for Non-stationary Sequence

In recent years, many methods[11-12] based on stochastic gradient descent have been proposed for neural network models. The gradient formula is:

$$g_t = \nabla_\theta L(f(x_t;\theta), y_t) \quad (1)$$

Among them, L is the loss function, and the most common one is the mean square error criterion. The model continuously updates the value of the parameter according to the given data sequence. The update formula of the basic stochastic gradient descent algorithm is as follows:

$$\theta_t = \theta_{t-1} - \alpha g_t \quad (2)$$

Unlike the previous variable update based on the data of the current iteration step, Adam updates the variables θ based on the gradient estimation of all data points before the current time step t. Adam's update rules are as follows:

$$\theta_t = \theta_{t-1} - \alpha \hat{m}_t / \hat{v}_t \quad (3)$$

The $\hat{m}_t = m_t(1-\beta_1^t)$ is the moment estimation version of m_t. It is the weighted sum of current and history variables g_t, which can be regarded as an estimate of expectations. The update rules are as follows:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\cdot g_t \quad (4)$$

Similarly, $\hat{v}_t = v_t(1-\beta_2^t)$ is the moment estimation version of v_t, which is the weighted sum of current and historical variables $g_t \odot g_t$. The update rules are as follows:

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\cdot g_t \odot g_t \quad (5)$$

When the sequence is a non-stationary sequence, the statistical variables change with time t, and it is unreasonable to use the gradient far from the current step to determine the update direction of the current step. Yulai Zhang [13] studied this type of problem and added a window variable k. Assume that in sequence Y, any sub-sequence of length $k \ll T$ has stable statistical variables, so it can be approximately regarded as a stationary sequence[14]. The renewed update rules are as follows:

$t \le k$:

$$\hat{m}_t = m_t(1-\beta_1^t) \quad (6)$$

$$\hat{v}_t = v_t(1-\beta_2^t) \quad (7)$$

$t > k$:

$$\overline{m}_t = m_t - \beta_1^k m_{t-k} \quad (8)$$

$$\overline{v}_t = v_t - \beta_2^k v_{t-k} \quad (9)$$

$$\hat{m}_t = \overline{m}_t(1-\beta_1^k) \quad (10)$$

$$\hat{v}_t = \overline{v}_t(1-\beta_2^k) \quad (11)$$

Finally:

$$\theta_t = \theta_{t-1} - \alpha \hat{m}_t / (\sqrt{\hat{v}_t}) \quad (12)$$

B. BiLSTM Optimized Based On NS-ADAM

The static window length k value in the NS-ADAM algorithm is artificially specified, and the choice of the k value plays a key role in the realization of the NS-ADAM algorithm. However, no scholar has stated how to find the best fixed window length method. This variable has a clear correlation with the data type itself, and it also has a significant impact on the accuracy of the prediction. In this paper, the general evaluation index MSE of the model is used to judge the algorithm under different window length k, and k-MSE is drawn into a graph. The lowest point of the curve can be designated as the optimal window length of the data set. In Figure 1, the mean square error under different window length k is given, and the model is carried out under the number of training times of 1200, and other parameters remain unchanged. When k is around 20, the fitting effect of the model is the best.

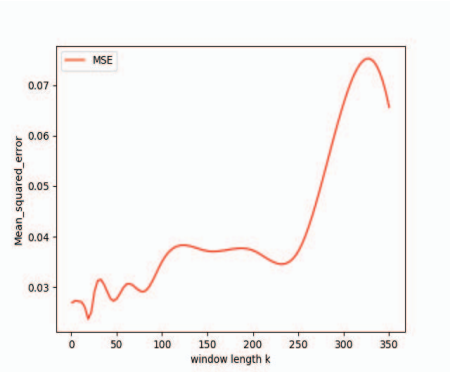


Figure 1 MSE of NS-ADAM algorithm

The NS-ADAM algorithm is based on the Adam algorithm and adds a static fixed window length k. At the same time, it proposes the assumption that items within k lengths can be approximately regarded as a stationary sequence, which enhances the correlation between the step direction of the gradient update and the previous gradient Sex. BiLSTM is an extended model of LSTM, which can be seen as a two-layer neural network. The first layer is from left to right as the input sequence of the model data sequence, and the second layer is from right to left as the sequence input sequence. This paper also uses the previously tested NS-ADAM optimization algorithm with a window length of 20 in the model to iteratively update the weights. The general structure is shown in Figure 2

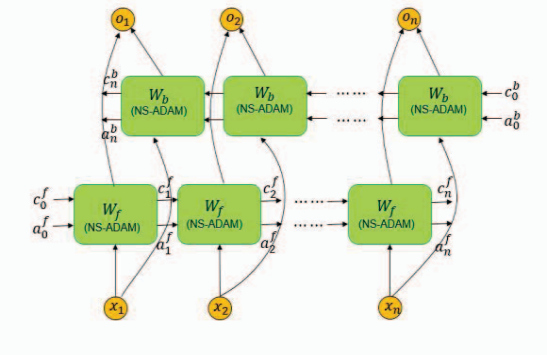


Figure 2 The new structure of BiLSTM

IV. COMPARATIVE ANALYSIS OF EXPERIMENTS

A. Data Introduction Preprocessing

This paper uses 46 key cities across the country as the data sampling area, and collects the order review data of some online catering merchants in the city's location, and the data is of the order of one million. After that, a risk assessment model was established with analytic hierarchy process and natural language methods as the main technical core, and the online food risk assessment value of each business was obtained. The author selects the risk value of a certain store from the final total data set as the data set for this article.

The original data is the food safety risk value of the merchant's network, and the data unit is generally between $10^{-2} \sim 10^{-4}$. In order to train the model better, the data is normalized and preprocessed. The normalization process maps the original data to the range [0,1]. The normalization formula[15] is shown in formula(13). x is the original data value, x' is the normalized data value.

$$x' = (x - x_{min}) / (x_{max} - x_{min}) \quad (13)$$

B. Model Evaluation Criteria

The loss value in the model is usually reported by the deep learning algorithm. Objectively speaking, the loss value is a punishment for poor prediction. If the model's prediction is completely accurate, then the loss value will be zero. Therefore, the model minimizes the loss as much as possible by obtaining a set of weights and deviations that minimize the loss value. In addition to the loss of optimization algorithms in deep learning, researchers often use the root mean square error (RMSE) as an indicator to judge the predictive performance of the model. The root mean square error represents the deviation of the observed value from the correct value. The formula is as follows:

$$RMSE(X,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)^2} \quad (14)$$

In the formula, m is the total number of samples, h is the predicted value, and y is the actual value. The main benefit of using RMSE is that it penalizes larger mistakes. In addition, we also use the percentage of RMSE reduction as an indicator to evaluate improvement. The calculation formula is as follows:

$$\%Reduction = \frac{New\_Value - Origina\_Value}{Original\_Value} \quad (15)$$

C. Comparison With Traditional Optimization Algorithms

This paper uses the BiLSTM neural network model to predict and analyze non-stationary sequences based on the food safety risk value data of the merchant network. The experiment compares the new optimization algorithm NS-ADAM and the traditional optimization algorithm[16-20] Adam, RMSprop, SGD under different batch sample numbers. Mean square error.

TABLE I. RMSE UNDER DIFFERENT ALGORITHMS UNDER BILSTM MODEL

Batch_size	SGD	RMSprop	Adam	Ns-ADAM
128	185.841	166.435	**78.644**	99.587
256	269.261	171.889	170.568	**142.208**
350	151.838	128.369	130.744	**89.552**
512	239.220	185.524	264.929	**167.801**
Average	211.540	163.054	161.221	**124.787**

TABLE II. COMPARISON WITH TRADITIONAL ALGORITHMS

Algorithm Comparison	%Reduction
NS-ADAM OVER SGD	-41.01
NS-ADAM OVER RMSPROP	-23.47
NS-ADAM OVER ADAM	-22.60

It can be seen from Table I of the experimental results that NS-ADAM has the smallest average root mean square error under the condition of the number of batch training samples from 128 to 512, and the best results are also obtained in most cases. The model has the best training effect when the number of batch samples is about 350. Table II reports the model improvement degree of NS-ADAM compared with the traditional optimization algorithm. Compared with SGD, the evaluation index of NS-ADAM has dropped by as much as 41.01%, and for the RMSprop and Adam algorithms, it has also dropped by more than 20%. Experiments prove that the BiLSTM model based on the NS-ADAM optimization algorithm significantly improves the prediction accuracy than the traditional BiLSTM model.

D. Comparison with traditional models

For the rigor of the experiment, this paper compares the BiLSTM model with the traditional LSTM network model and the integrated moving average autoregressive model ARIMA model. In this paper, the original data set is normalized and preprocessed, and the processed data is input into the ARIMA, LSTM, and BiLSTM models as the only feature of the risk assessment sequence. The data set is divided into two parts: training and testing. 90% of the data set is used for training and 10% of the data set is used to test the accuracy of the model. Table III shows the model comparison of various optimization algorithms under the LSTM model and the comparison with the BiLSTM and ARIMA models.

TABLE III. RMSE OF ARIMA, LSTM AND BILSTM MODELS

Model	RMSE
ARIMA	345.262
LSTM_SGD	205.684
LSTM_RMSPROP	168.925
LSTM_ADAM	166.978
LSTM_NSADAM	153.852
BILSTM_NSADAM	89.5524

Table III reports the root mean square error achieved by each model that predicts the risk value of online food safety for merchants. It can be clearly seen that the LSTMs-based recurrent neural network model has a significantly lower RMSE value than the traditional autoregressive model ARIMA. In the one-way LSTM model, the RMSE value of the LSTM using the NS-ADAM optimization algorithm is also smaller than that of the one-way LSTM model using other optimization algorithms. And it can be seen that the performance of the BiLSTM model is better than the one-way LSTM model, and the gap is very large. In summary, the performance of the BiLSTM model is the best among the three types of models, and the BiLSTM model based on the NS-ADAM optimization algorithm performs best under various optimization algorithms. The experiment also compares the training time of BiLSTM and LSTM models. The average training time of the BiLSTM model is about 700s, and the average training time of the LSTM model is about 470s. Although the accuracy of the BiLSTM model is improved compared to the LSTM model, the time performance is reduced. Given that the internal structure of the BiLSTM model is more complex than that of the LSTM model, this experimental result is reasonable. Figures 3, 4, and 5 are the prediction diagrams of the food safety risk values of the three models.

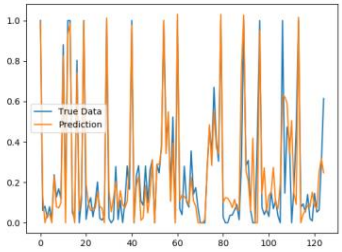


Figure 3 BiLSTM

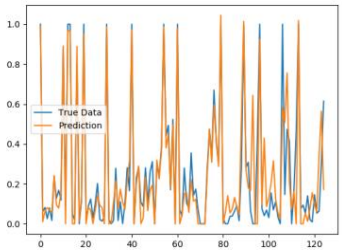


Figure 4 LSTM

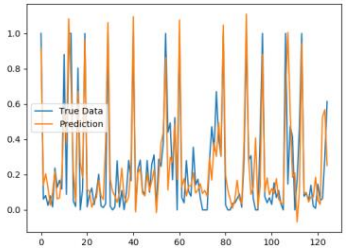


Figure 5 ARIMA

V. SUMMARY AND OUTLOOK

This article reports the results of an experiment in which the performance and accuracy of various optimization algorithms based on the BiLSTM model are compared and analyzed in the online food safety risk data prediction experiment. At the same time, analysis and comparison between different models of ARIMA, one-way LSTM and BiLSTM are carried out. The research question for this experiment focuses on the prediction of non-stationary series. Does the application of the NS-ADAM optimization algorithm in the LSTMs model have any positive impact on improving the accuracy of the time series? The results show that the application of the NS-ADAM algorithm helps the model improve the accuracy of prediction, and the effect is significant, which is conducive to modeling. We noticed that the NS-ADAM algorithm has significant advantages in either the one-way LSTM or the BiLSTM model. Compared with the LSTM model, the accuracy of the BiLSTM model is improved, but the training speed is relatively slow. When time conditions permit, this article recommends using the BiLSTM model combined with the NS-ADAM algorithm to predict non-stationary series. This research can be further extended to forecasting problems for multivariate and seasonal time series.

REFERENCES

[1] Zhang Shukai,Lei Xin. Internet Takeaway Food Safety Regulatory lssues[J].MODERN FOOD, 2016, 000(003):P.41-45.

[2] Wahyuni H C, Vanany I, Ciptomulyono U. Application of Bayesian Network for Food Safety Risk in Cattle Slaugtering Industry[C].2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). IEEE, 2019: 450-454.

[3] Ji Lihong, Chen hong. Problems in the supervision of online food delivery industry and their governance [J]. WESTERN SO, 2019(31)

[4] Cai Yuanzheng. The Dilemma and Outlet of Online Food Safety Supervision [J]. LEGALITY VISION, 2017 (3): 11.

[5] Kim J , Moon N . BiLSTM model based on multivariate time series data in multiple field for forecasting trading area[J]. Journal of Ambient Intelligence and Humanized Computing, 2019(5).

[6] Cui Z , Ke R , Pu Z , et al. Deep Bidirectional and Unidirectional LSTM Recurrent Neural Network for Network-wide Traffic Speed Prediction[J]. 2018.

[7] Lee S I , Yoo S J . Threshold-Based Portfolio: The Role of the Threshold and Its Applications[J]. Papers, 2018.

[8] Krauss, Christopher, Anh Do, Xuan, Huck, Nicolas. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500[J]. European Journal of Operational Research, 2016:S0377221716308657.

[9] De Baets L, Ruyssinck J, Peiffer T, et al. Positive blood culture detection in time series data using a BiLSTM network[J]. arXiv preprint arXiv:1612.00962, 2016.

[10] Sha S, Li J, Zhang K, et al. RNN-Based Subway Passenger Flow Rolling Prediction[J]. IEEE Access, 2020, 8: 15232-15240.

[11] Liu J, Qian L, Zhang Y, et al. Towards Safety-Risk Prediction of CBTC Systems With Deep Learning and Formal Methods[J]. IEEE Access, 2020, 8: 16618-16626.

[12] Siami-Namini S, Tavakoli N, Namin A S. The performance of LSTM and BiLSTM in forecasting time series[C]//2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019: 3285-3292.

[13] Bukhari A H, Raja M A Z, Sulaiman M, et al. Fractional neuro-sequential ARFIMA-LSTM for financial market forecasting[J]. IEEE Access, 2020, 8: 71326-71338.

[14] Zhang Y, Wang Y, Luo G. A new optimization algorithm for non-stationary time series prediction based on recurrent neural networks[J]. Future Generation Computer Systems, 2020, 102: 738-745.

[15] Zhang Z. Improved Adam optimizer for deep neural networks[C].2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). IEEE, 2018: 1-2.

[16] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[17] Liu H, Xu H, Yan Y, et al. Bus Arrival Time Prediction Based on LSTM and Spatial-Temporal Feature Vector[J]. IEEE Access, 2020, 8: 11917-11929.

[18] Kalchbrenner N, Danihelka I, Graves A. Grid long short-term memory[J]. arXiv preprint arXiv:1507.01526, 2015.

[19] Zhao Y. Spatial-Temporal Correlation-Based LSTM Algorithm and Its Application in PM2. 5 Prediction Spatial-Temporal Correlation-Based LSTM Algorithm and Its Application in PM2. 5 Prediction[J].

[20] Li Y, Bao T, Gong J, et al. The Prediction of Dam Displacement Time Series Using STL, Extra-Trees, and Stacked LSTM Neural Network[J]. IEEE Access, 2020, 8: 94440-94452.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Coefficient Optimization Based on Genetic Algorithm for Hybrid Control of Semi-active Suspension System

Li Wei
School of Automotive Engineering
Wuhan University of Technology
Wuhan, China
1490239828@qq.com

Zhang Chengcai
School of Automotive Engineering
Wuhan University of Technology
Wuhan, China
zhangchc@whut.edu.cn

***Abstract*—Composed of sky-hook control and ground-hook control, hybrid control used for a semi-active suspension system is of great significance for enhancing vehicle comfort, stability, and safety. The former hybrid control is highly difficult in coefficient selection to lead to has a low degree of improvement in suspension performance. In this paper, based on a genetic algorithm, a coefficient optimization method (COGA) is presented for hybrid control of a semi-active suspension system. A genetic algorithm is used to optimize the selection of suspension control system coefficients. Computer simulation of the quarter-vehicle suspension system model is conducted by MATLAB/SIMULINK software. The vehicle comfort and handling characteristics of the suspension systems are predicted under a random road input simulation model. Passive suspension system performance is compared with the semi-active suspension system using hybrid control. The experimental results show that the semi-active suspension hybrid control system with COGA has a minimal root-mean-square value of sprung mass acceleration, suspension dynamic deflection, and dynamic tire load. It can be proved that COGA can effectively coordinate and ameliorate the function of hybrid control.**

Keywords-hybrid control; semi-active suspension system; genetic algorithm; coefficient optimization

I. INTRODUCTION

The vehicle suspension system plays an important role in a vehicle by providing vehicular stability and comfort during driving by isolating its passengers and goods for vibration caused by the road surface as a result of irregularities like potholes, speed bumps, and uneven road surfaces. Nowadays, vehicle suspension may be categorized broadly into three types: passive suspension, active suspension, and semi-active suspension. Due to its constant damping coefficients, the passive suspension system leads to a compromise between ride and handling, and can't adapt to a changing road. Both active suspension systems and semi-active suspension systems are capable of changing their damping characteristics by using an amount of external power. The active suspension system has the disadvantages of costliness, complexity, and high energy consumption, there are drawbacks of which that stop it from being extensively applied in most vehicles. Unlike an active suspension system, the semi-active suspension system is simpler, cheaper, and more reliable, which is becoming more and more popular for car manufacturers.

In 1974, Crosby et al. proposed a shy-hook control and applied it to a passive suspension system, resulting in a semi-active suspension system [1]. With the advancement of technology, there have been many control strategies for the semi-active suspension system. Such as, ground-hook control [2], PID control [3], sliding mode control [4], fuzzy control [5], optimum control [6], and neural network control [7]. However, the vehicle performance contradiction between riding comfort and handling stability under the single working model of the semi-active suspension system has always existed. For example, sky-hook control significantly improves vehicle riding comfort but doesn't work on handling stability. Ground-hook control primarily improves vehicle handling stability but will deteriorate riding comfort [8].

Ahmadian et al. put forward a hybrid control based on shy-hook control and ground-hook control to coordinate the contradiction between comfort and stability of the semi-active suspension system [9]. Hybrid control is simple, easy to implement, and has a more excellent control effect than the single control strategy, so it is widely used in vehicles. At the same time, hybrid control is often inefficient on account of the wrong choice of damping coefficients. Researchers have used a particle swarm optimization algorithm (PSO) to optimize the damping coefficients of hybrid control [10]. But the result is not optimal and can be further optimized by increasing the weight coefficient optimization. The genetic algorithm is a global search heuristic algorithm that is proposed by Holland et al [11-12]. There have been some control strategies with a genetic algorithm to optimize control system parameters to get outstanding suspension performance. As a consequence, this paper proposes a coefficient optimization method (COGA) for hybrid control of the semi-active suspension, which uses a genetic algorithm to optimize its damping coefficients and weighting coefficient.

II. MODEL OF SEMI-ACTIVE SUSPENSION SYSTEM UNDER HYBRID CONTROL

Sky-hook control realizes damping control by installing a damping element between the sprung mass and the imaginary inertial reference frame. In the same, ground-hook control realizes damping control by installing a damping element between the unsprung mass and the imaginary inertial reference frame. In this paper, a quarter vehicle dynamic suspension system model under hybrid control is established, as is shown in Figure 1. According to Newton's laws of motion, the dynamic motion equations for the quarter vehicle suspension can be expressed as:

$$\begin{cases} M_1\ddot{x}_1 = -k_0(x_1 - x_2) - c_0(\dot{x}_1 - \dot{x}_2) + F \\ M_2\ddot{x}_2 = k_0(x_1 - x_2) + c_0(\dot{x}_1 - \dot{x}_2) - k_t(x_2 - x_r) - F \end{cases} \quad (1)$$

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00053

where M_1 is the sprung mass, M_2 is the unsprung mass, k_0 is the spring stiffness, k_t is the tire stiffness, x_r is the displacement of road input, x_1 is the displacement of the sprung mass, x_2 is the displacement of the unsprung mass, c_0 is the passive damping coefficient, c_s is the sky-hook damping coefficient, c_g is the ground-hook damping coefficient, and F is the hybrid damping force.

The control law of hybrid strategy can be stated as:

$$F = -\alpha c_s \dot{x}_1 + (1-\alpha) c_g \dot{x}_2 \tag{2}$$

When α is equal to 1, hybrid control is pure sky-hook control:

$$F = -c_s \dot{x}_1 \tag{3}$$

When α is equal to 0, hybrid control is pure ground-hook control:

$$F = c_g \dot{x}_2 \tag{4}$$

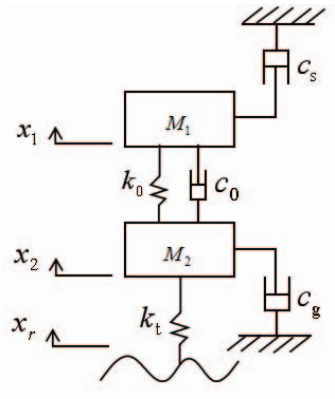


Figure 1. Suspension dynamic model with 2-DOF of hybrid control.

III. Coefficient Optimization Based on Genetic Algorithm

A genetic algorithm is a random adaptive global search method used for rapidly calculating optimization problems of multiple parameters, which is strong fault-tolerant, easy to implement, and has high computational efficiency. The genetic algorithm has been the focus of attention of researchers since it was proposed. Combining the genetic algorithm with other subjects has become the latest research trend.

Therefore, it is practicable to use a genetic algorithm to optimize suspension system control coefficients. Figure 2 shows the overall schematic diagram of COGA, which can be defined in the following steps:

1) Algorithm initialization. Initialize the entire population with completely random solutions.

2) Coefficient assignment. Respectively assign the population individuals to the sky-hook damping coefficient, the ground-hook damping coefficient, and the weighting coefficient.

3) Evaluate the fitness function. Calculate the fitness value for each member of the population. The fitness function L is constructed by suspension system performance evaluation indexes. So this paper uses the root-mean-square value of sprung mass acceleration, suspension dynamic deflection, and tire dynamic load to construct function, which formula is as follows:

$$\min L = 0.44\frac{SMA(x)}{SMA_p} + 0.28\frac{SDD(x)}{SDD_p} + 0.28\frac{DTL(x)}{DTL_p} \tag{5}$$

$$s.t. = \begin{cases} \frac{SMA(x)}{SMA_p} < 1 \\ \frac{SDD(x)}{SDD_p} < 1 \\ \frac{DTL(x)}{DTL_p} < 1 \end{cases} \tag{6}$$

Where SMA(x), SDD(x), DTL(x) is the root-mean-square value of sprung mass acceleration, suspension dynamic deflection, and tire dynamic load of the semi-active suspension, and SMA_p, SDD_p, DTL_p is the root-mean-square value of sprung mass acceleration, suspension dynamic deflection, and tire dynamic load of the passive suspension. In order to greatly improve vehicle riding comfort, the ratio of sprung mass acceleration is increased.

4) Termination condition. The maximum number of iteration is used as a termination condition. If the termination condition is met, output the optimization coefficients. if not, step five is executed.

5) Selection, crossover, and mutation. Apply mutation, crossover, and selection operators to the individuals with the smallest fitness value to generate a new population. Repeat steps two to four.

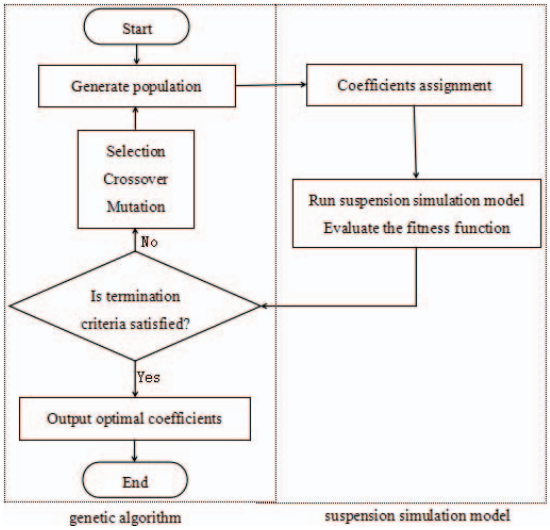


Figure 2. The process of coefficient optimization method based on genetic algorithm (COGA).

IV. Experiment

A quarter semi-active suspension hybrid control system simulation model is established in MATLAB/SIMULINK, as shown in Figure 3. As shown in Figure 4, a random road input simulation model is established by a filtered white noise signal [13], which formula is as follows:

$$\dot{x}_r(t) = -2\pi f_0 x_r(t) + 2\pi n_0 \sqrt{G_q(n_0)u}\, w(t) \tag{7}$$

where $G_q(n_0)$ is the road irregularity coefficient, f_0 is the lower cutoff frequency, u is the vehicle speed, w(t) is the unit white noise, n_0 is the reference space-frequency, and $x_r(t)$ is the displacement of road input.

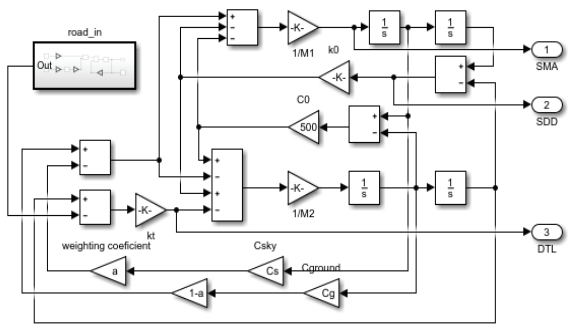


Figure 3. Semi-active suspension system simulation model.

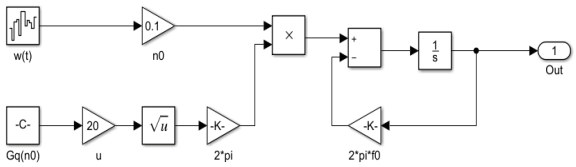


Figure 4. Random road input simulation model.

According to the parameters in Table Ⅰ, the simulation environment is set. COGA is carried out in MATLAB/SIMULINK. The value range of the sky-hook damping coefficient is [0, 10000]. The value range of the ground-hook damping coefficient is [0, 10000]. The value range of the weighting coefficient is [0.4, 0.6]. As shown in Figure 5, the coefficient optimization results are listed as follows: c_s = 4460.9Ns/m, c_g= 604.7954Ns/m, α = 0.5809.

TABLE I. SIMULATION PARAMETERS

Parameter	Symbol	Values
Sprung mass (Kg)	M_1	250
Unsprung mass (Kg)	M_2	45
Spring stiffness (N/m)	k_0	16000
Tire stiffness (N/m)	k_t	160000
Passive damping coefficient (Ns/m)	c_0	500
Vehicle speed (m/s)	u	20
Road irregularity coefficient (m^3)	$G_q(n_0)$	256e-6
Lower cutoff frequency (Hz)	f_0	0.01
Reference space-frequency (m^{-1})	n_0	0.1
Sky-hook damping coefficient (Ns/m)	c_s	[0, 10000]
Ground-hook damping coefficient (Ns/m)	c_g	[0, 10000]
Weighting coefficient	α	[0.4, 0.6]
Population size	-	100
Elite count	-	10
Crossover fraction	-	0.8
Generation	-	20

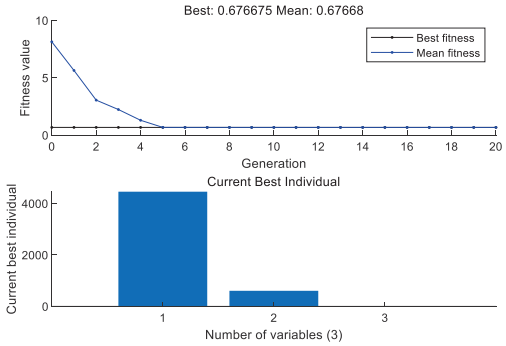


Figure 5. Result of coefficients optimization.

Figure 6 shows the response of the passive suspension system, the semi-active suspension hybrid control system with a particle swarm optimization algorithm (PSO-hybrid control), and the semi-active suspension hybrid control system with COGA (COGA-hybrid control) under the random road input simulation model. It can be seen from the image that compared with the passive suspension system, the semi-active suspension system can drastically reduce the amplitude of suspension system performance evaluation indexes. Most importantly, the semi-active suspension hybrid control system with COGA has the smallest response peak. As can be roughly seen from the figure, COGA can effectively improve the efficiency of hybrid control more than PSO.

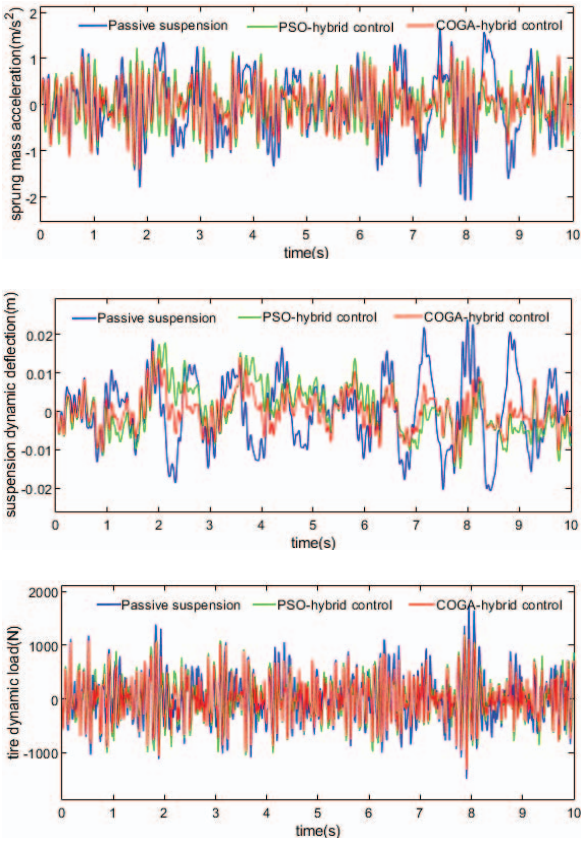


Figure 6. Responses of suspension evaluation indexes.

To further analyze the performance of the suspension systems, the root-mean-square value of evaluation indexes is calculated, and the control effect of the suspension systems is evaluated, as shown in Table Ⅱ. Table Ⅱ shows that the three indexes of the hybrid control suspension system are reduced with different degrees under the random road input simulation model. For example, under the random road input simulation model, the root-mean-square value of the sprung mass acceleration of hybrid control with COGA is reduced by 32.2% compared with that of the passive suspension. The root-mean-square value of the sprung mass acceleration of hybrid control with PSO is reduced by 29.1% compared with that of the passive suspension. It can be seen that the root-mean-square value of the sprung mass acceleration of hybrid control with COGA goes down more than hybrid control with PSO. The same is true for suspension dynamic deflection and tire dynamic load. It is obvious that the control effect of the hybrid control with COGA is better than the hybrid control with PSO.

TABLE II. THE OUTPUT RMS VALUES AND CONTROL EFFECT

control types	Passive suspension	PSO-hybrid control	COGA-hybrid control
SMA	0.6805	0.4827	0.4613
Effect(%)	-	29.1	32.2
SDD	0.009015	0.006115	0.004672
Effect(%)	-	32.1	48.2
DTL	462.6	405.8	386.9
Effect(%)	-	12.3	16.4

V. CONCLUDING REMARKS

A coefficient optimization method (COGA) is designed, based on a genetic algorithm, which is used to a hybrid control strategy of the semi-active suspension system. In this work, a genetic algorithm is serviced to optimize the sky-hook damping coefficient, the ground-hook damping coefficient, and the weighting coefficient of hybrid control. The simulation analysis of the suspension systems is carried out by MATLAB/SIMULINK software. Compared with the passive suspension system and the semi-active suspension hybrid control system with PSO, the root-mean-square value of sprung mass acceleration, suspension dynamic deflection, and tire dynamic load of the semi-active suspension hybrid control system with COGA has different degrees of reduction. The experimental results show that hybrid control with COGA has the best control effect, which can improve the vehicle riding comfort and handling stability, and prevent the adverse effects of body resonance on human organs effectively. It is proved that COGA can effectively improve the efficiency of hybrid control.

ACKNOWLEDGMENT

This work was supported by the Independent Innovation Research Fund Project of Wuhan University of Technology (NO. 2021-QC-B1-11).

REFERENCES

[1] Karnopp D. Design principles for vibration control system using semi-active damper[J]. Journal of Dynamic Systems, Measurement and Control, 1990. 112(3): 448-55.

[2] Val S.M, Noxak M, Sika Z, et al. Extended Ground-Hook-New Concept of Semi-Active Control of Truck's Suspension[J]. Vehicle System Dynamics, 1997, 273(5-6): 289-303.

[3] Yang Jianwei, Li Jie, Jia Zhixuan, et al. Fuzzy-PID Country System Simulation of the Semi-active Vehicle Suspension. 2010. (1): 772-775.

[4] Zhiyong Yang et al. Sliding Mode Control for Vibration Comfort Improvement of a 7-DOF Nonlinear Active Vehicle Suspension Model[J]. Journal of Robotics and Mechatronics, 2019, 31(1): 95-103.

[5] H Pang, W Fu, K Liu. Stability analysis and fuzzy smith compensation control for semi-active suspension systems with time delay[J]. Journal of Intelligent & Fuzzy systems, 2015. 29(6): 2513-2525.

[6] Garg A, Arvind A, Gahvi B. Optimum Control for the Vehicle Semi-active Suspension System[C]. 2017: 421-430.

[7] Zhushun Ding, Feng Zhan, Yechen Qin, et al. Adaptive Neural Network Control for Semi-active Vehicle Suspensions[J]. Journal of Vibroengineering, 2017, 19(4): 2654-2669.

[8] Jalwadi, M.S. and Unaune, D. Performance analysis of Skyhook, groundhook and hybrid control strategies on semi-active suspension system. International Journal of Current Engineering and technology, Special Issue 3, 2014: 265-269.

[9] Ahmadian M, Pare C A. A Quarter-Car Experimental Analysis or Alternative Semiactive Control Methods[J]. Journal of Intelligent Material Systems and Structures, 2000, 11(80): 604-612.

[10] WU Wei-chun, CHEN Long, DING Ren-kai. Research on Hybrid Control Strategy of Active Suspension Based on Particle Swarm Algorithm [J]. Machinery Design & Manufacture, 2018(01): 29-31+35.

[11] Holland J H. Adaptation in Nature and Artificial Systems[J]. 2nd ed Cambridge: MIT press, 1992: 91-92.

[12] Holland J H. Genetic algorithms[J]. Scientific American, 1992, 267(1): 66-72.

[13] GUO Shanglai. Random control [M]. Tsinghua University Press, 1999: 97-117.

Research on Historical Concept Modeling Language and Modeling Tool based on Model Driven Architecture

Mingzhu Ma,Hongxing Liu
College of Computer and Artificial Intelligence
Wuhan University of Technology
Wuhan, China
1464419897@qq.com

Zhenqi Li
Server R&D
Wuhan Baijia Cloud Technology Co., Ltd.
Wuhan, China
782371833@qq.com

***Abstract*—This paper proposes a framework for a model-driven historical compilation method. The basic idea is to distinguish historical conceptual models and historical works models, so as to support large and medium-scale historical compilations of "top-down, gradual refinement". It focuses on defining the historical concept modeling language, and designing its abstract grammar, concrete grammar and semantics. Based on MetaEdit+, a preliminary graphical historical concept modeling tool is designed and implemented.**

Keywords-component; model-driven; historical concept modeling language; historical concept modeling tool

I. INTRODUCTION

The compilation of historical works is referred to as historical compilation. In the traditional compilation practice, after collecting certain historical materials and documents, historians integrate relevant historical materials and documents, and finally describe history through words. The problems brought about by this compilation process are the lack of a unified and recognized compilation process and method, the lack of detailed definition and division of the compilation process, and the lack of high-level professional auxiliary tools. This article considers the application of modeling ideas in the computer field to the field of historical compilation to assist historians in the compilation of history. In this process, through the historical modeling, the work results of each stage of historical compilation are formally expressed.

This paper attempts to explore the method of computer-assisted compilation of historical works from a higher level. It believes that history is essentially a number of historical models used to describe related events, people, environments, objects and connections in a specific world. Historical model is a representation of history, historical compilation is historical modeling, and the language used in historical modeling is called historical modeling language. Drawing on the successful experience in the field of software and databases, we believe that the compilation of historical works can distinguish conceptual models, logical models, and physical models from abstract to concrete. The historical conceptual model is a basic platform-independent model that can be converted to a variety of platform-related models. In this paper, we focus on the conceptual model of history.

In the second II, we propose a model-driven history compilation method. Section III define the history concept modeling language. In the section IV, we implement a graphical history modeling tool. Section V summarizes the research content of this paper.

II. MODEL-DRIVEN HISTORICAL COMPILATION METHOD

A. The basic idea of model-driven architecture

Model-driven architecture (MDA) is a new software development method proposed by the Object Management Organization[1][2]. The basic idea of MDA is to separate the model from the implementation, that is, to separate the expression of the model from the implementation of the model by taking the model as the center. MDA proposed the concepts of Platform Independent Model (PIM) and Platform Specific Model (PSM). Platform refers to a series of resources needed to realize the system. For example, in the field of historical compilation, the technology and carrier of historical compilation can be regarded as platforms. PIM has nothing to do with specific platforms and is a highly abstract model. The PSM is related to the specific platform and is a less abstract model. Under the framework of MDA, the process of software development is that domain experts model the business logic of the system to form PIM, then automatically or semi-automatically convert the PIM to the PSM under each platform, and finally, the source code is generated by the PSM[3][4].

B. Model-driven historical compilation method

MDA software design idea is introduced into the field of historical compilation to explore the model-driven historical compilation method. The main content of the model-driven historical compilation method is to design and establish historical models, that is, historical modeling. Complete historical modeling includes historical conceptual modeling, historical logic modeling, and historical physical modeling.

The architecture of the model-driven historical compilation method can be represented in Figure 1.

From the perspective of MDA modeling, historical works is essentially a model, which we call historical models. Historical works belong to narrative works. Using narratology theory[5] to analyze historical works, historical works can be divided into "story layer" and "discourse layer". According to the abstract level of modeling, MDA distinguishes platform independent model (PIM) and platform specific model (PSM). Regarding the target

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00054

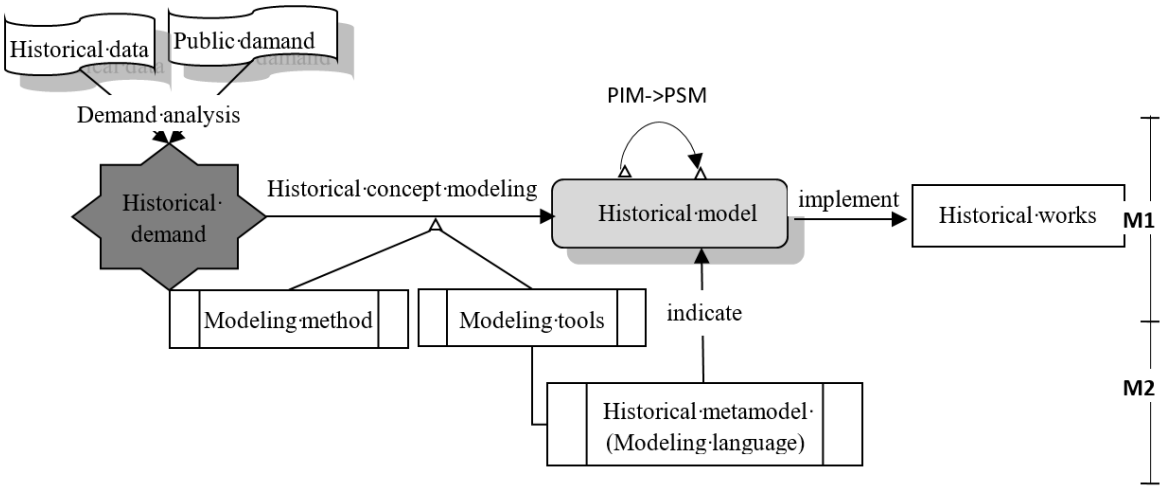


Figure 1. Framework for a model-driven historiography approach

media or carriers of story works (such as books, movies, animations, etc.) as a platform for carrying stories, then the "story layer" of historical works can be regarded as PIM, "discourse layer" " can be regarded as PSM. MDA divides the model into 4 layers. We focus on the model layer (M1) and the meta model layer (M2).

The historical conceptual model is a PIM. When modeling historical concepts, according to the "God's perspective", arrange events based on the actual time when the event occurred, and describe the connection between the events and the connection between the events and the main characters and places. At this time, the rhetoric is not considered. On the basis of historical conceptual models, consider the use of various "discourse" methods, focusing on narrative perspective (such as the third person), narrative time (such as flashback), narrative space and other methods, and finally form historical works (PSM). The model-driven history compilation method fully draws on the experience of database design and other fields, considers the separation of concerns, and the use of auxiliary tools, which helps historians to compile large and medium-sized historical works.

III. HISTORICAL CONCEPT MODELING LANGUAGE

The historical concept modeling language is a modeling language used to express historical concept models. This section describes the historical concept modeling language from three aspects: abstract syntax, concrete syntax and semantics.

A. Abstract syntax

Abstract syntax defines what types of components exist in the modeling language, and how these components are combined by other components [6]. In this paper, abstract syntax is composed of building blocks and general rules.

This paper uses MetaEdit+[7] meta-model GOPPRR to design building blocks. The building blocks in the historical concept modeling language are designed as follows:

1) Design of building block Graph: Create a Graph from Graph Tool and name it "History Modeling Tool". "History Modeling Tool" contains two attributes: Model Name and Model Description, which are used to record the model name and the description of the model.

2) The design of the building block Object: The "event" and other elements in the meta-model are designed through Object. There are seven types of Object type elements, as shown in Table I.

TABLE I. ELEMENT OF TYPE OBJECT

Object	***Attributes***	***instructions***
Start	no	The beginning of the model
End	yes	End of model, including model end and process end
Event	yes	Major elements in the historical model
Character	yes	The person concerned with the event
Context	yes	Time, place, etc
Article	yes	An important tool, device, invention, etc
Timeline	yes	The date of the incident

3) The design of the building block Property: each property contains the property name and property type.The representation of the attribute is as show in formula(1).

$$NP.Property::Type \quad (1)$$

NP represents the non-attribute building block in the meta-model, Property represents the attribute that NP has, and Type represents the type of the Property.

B. Concrete syntax

In this paper, the concrete syntax is the graphical representation of the modeling element. There is a close connection between the graphical representation of the modeling element and the abstract syntax, which is usually defined by mapping the abstract syntax to the notation[8].Figure 2 shows the mapping between abstract syntax and concrete syntax.

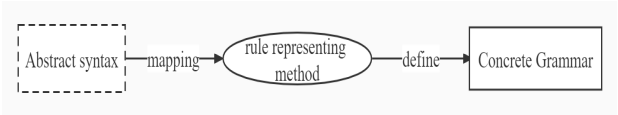


Figure 2. From abstract syntax to concrete syntax

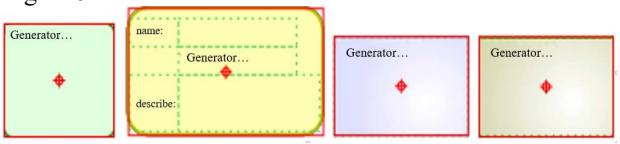
With the help of graphical editing tool "Symbol Editor", specific grammar can be easily realized. The modeling elements of the graphical representation designed in this paper include three types:

1) Object elements, including Event, Character, Context and Article.

2) Contact elements, including Flow, External Flow, Link, and Projection.

3) Role elements including From, To, External Flow From, External Flow To, Link From, Link To, Projection From, and Projection To.

The concrete syntax is shown in Figure 3, Figure 4 and Figure 5.



(a)event (b)Character (c)Context (d)Article

Figure 3. A concrete representation of object

(a) External Flow From (b) Flow From (c) Link From

d) External Flow To (e) Flow To (f) Link To

Figure 4. A concrete representation of a role

(a) External Flow (b) Flow (c) Link

Figure 5. A concrete representation of the relationship icon

C. *The semantic*

Semantics is the interpretation and logical representation of the meta-model[9][10]. In the paper, the semantics is designed for the field of historical modeling, it is the explanation and description of the historical meta-model. If it is separated from the field of historical conceptual modeling, the semantics is meaningless. In the design of this article, the semantics are mainly restricted by the model interpreter

The model parser is mainly the editor generator definition language MERL. Through the editing of the MERL code, the model can be detected in real time, errors in the model can be found in time, and corresponding prompt information can be given according to actual needs.

As shown in Figure 6, in the model parser, two parts are defined. In the upper rectangular box are the corresponding inspection rules designed. Click each line to display the corresponding MERL code in the lower rectangular box.

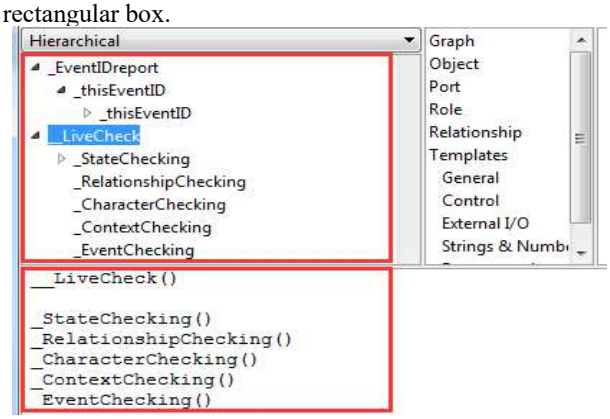


Figure 6. Model parser interface

IV. DEVELOPMENT OF HISTORICAL CONCEPT MODELING TOOLS

The tool interface includes:

- The layout of object elements and associated elements in the toolbar.
- Drawing area of model graph.
- location of information indicating violation of integrity constraints.
- Display position of elements in the model diagram.

The modeling tool should be able to visually display the connections between events, characters, environments and objects. Historians can create and delete elements of events, characters, environments and objects through historical modeling tools, and be able to edit the attributes of these elements. Two elements can be connected through some kind of connection. When performing these operations, the historical modeling tool should be able to check the integrity constraints in real time. If the integrity constraints are violated, information prompts will be given to facilitate the modification of the model. The interface of the historical modeling tool is shown in Figure 7.

V. CONCLUSION

This article explores the model-driven historical compilation method, focusing on the historical concept modeling language. It defines the historical concept modeling language from three methods: abstract syntax, concrete syntax and semantics, and represents historical concept models from a higher level of abstraction. Finally, a preliminary historical modeling tool is implemented to support historical concept modeling. The research in this article is only a preliminary exploration of the model-driven history compilation method. Therefore, there are still many deficiencies in the design and implementation of the research and modeling tools. The basic syntax and integrity constraints of the historical modeling language need to be further improved. It can adapt to a more general history.

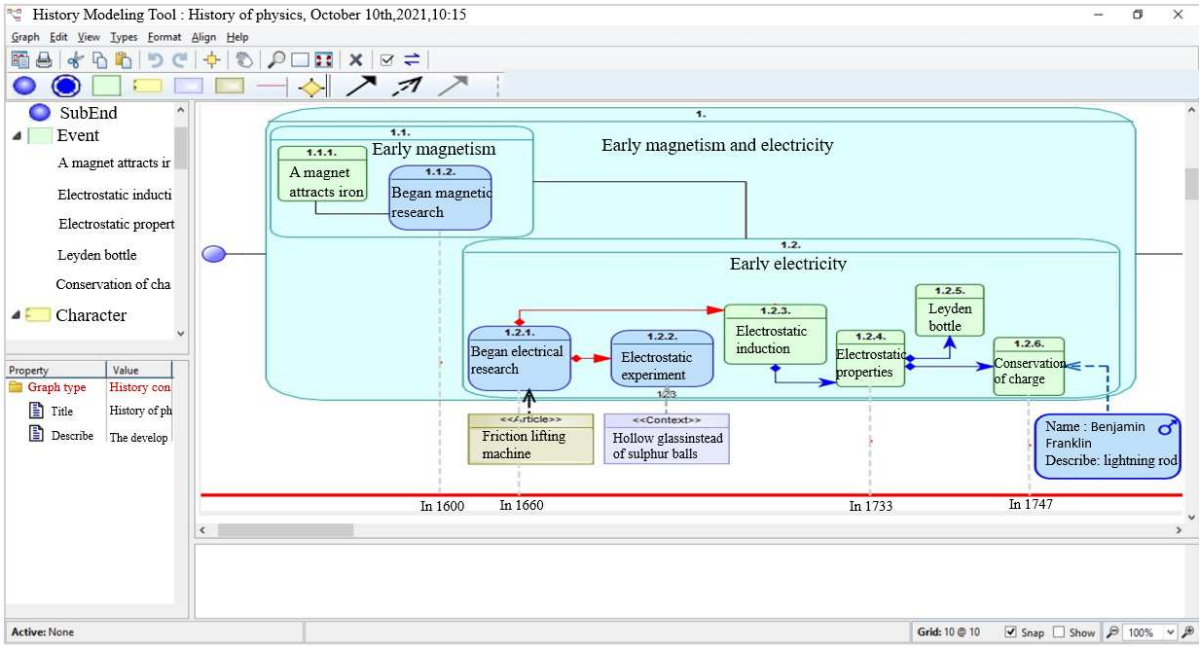


Figure 7. History modeling tool interface

REFERENCES

[1] OMG. Model Driven Architecture(MDA) Guide rev.2.0[EB/OL]. [2018-01]. http://www.omg.org/cgi-bin/doc?ormsc/14-06-01.

[2] OMG, THE OMG SPECIFICATIONS CATALOG [EB/OL]. [2018-01], https://www.omg.org/spec/.

[3] Hongxu Sun. Research and Realization of MDA Model Conversion Method[D]. Harbin :arbin Engineering University,2012.

[4] Yuanqing Zhang. Design and Implementation of Model Conversion Tool Based on MDA[D]. Harbin :arbin Engineering University,2011.

[5] Yinde Zhang, Narrative Research[C]. China Social Sciences Press,1986.

[6] Wile D S, “Abstract syntax from concrete syntax”, Proceedings of the 19th international conference on Software engineering. ACM, 1997,pp. 472–480.

[7] J.-P. Tolvanen, "MetaEdit+ for collaborative language engineering and language use (tool demo)",Proceedings of the 2016 ACM SIGPLAN International Conference on Software Language Engineering, 2016, pp. 41-45.

[8] D. Bork, D. Karagiannis and B. Pittl, "Systematic analysis and evaluation of visual conceptual modeling language notations", 2018 12th International Conference on Research Challenges in Information Science (RCIS),2018,pp.1-11.

[9] D. Durisic, C. Motta, M. Staron and M. Tichy, "Co-Evolution of Meta-Modeling Syntax and Informal Semantics in Domain-Specific Modeling Environments — A Case Study of AUTOSAR", 2017 ACM/IEEE 20th International Conference on Model Driven Engineering Languages and Systems (MODELS), 2017, pp. 189-198.

[10] Dániel Urbán, Zoltán Theisz and Gergely Mezei,” Self-describing operations for multi-level meta-modeling”, 6th Internationalh hokgk f Conference on Model-Driven Engineering and Software Development, 2018, pp. 519–527.

Fungal Growth Kinetics Model Based on Carbon Cycle

NI Aichen
Wuhan University of Technology, School of Economics
WUT
Wuhan,China
1751339693@qq.com

HUANG Qiang，CHENG Chuanbin
Wuhan University of Technology, School of Science
WUT
Wuhan,China
Huangh6h6@163.com

***Abstract*—Fungi play a critical role in the entire carbon cycle. In this article, the process of fungi participating in the carbon cycle was analyzed with differential equations, and a fungal growth kinetic model was set up through mechanism analysis. The results reflect that the optimum temperature required for different strains to reach the stable point is completely different, and the advantages and disadvantages of the growth of each strain under different temperature conditions are different. The model reveals the process of decomposing ground litter and wood fiber through its life activities in the presence of fungi.**

Keywords-Fungi;Carbon cycle;Growth kinetic model

I. INTRODUCTION

Fungus is a kind of eukaryote, which mainly uses organic matter produced by other organisms as a carbon source for proliferation. Because of this, it plays an irreplaceable role in both the local carbon flow and the overall carbon cycle [1] .

Existing research has been able to determine the relationship between the fungal decomposition rate of litter and wood fiber on the ground and some of its characteristics, and can quantify it through experimental data [2] ; however, research on the dynamic characteristics of fungi decomposing organic matter and multi-species interaction mechanism [3] is lacking. The former problem will be effectively solved in this article.

If a single fungal colony is simplified into a population, and a population growth model (usually a logistic model) is established on this basis, it can indeed reflect the process of fungi decomposing organic matter relatively effectively However, this type of model assumes the simplest linea relationship between population growth rate and substrate decomposition rate, and it is difficult to accurately describe the dynamic characteristics of the decompositior process [4] . So a different approach is taken which uses the idea of carbon cycle to describe the fungus's decomposition process of ground litter and wood fibe through the flow of carbon element. and thus the funga growth kinetic model can be built up.

The process of carbon transfer from the ground litter to the fungus for metabolism can be summarized as follows: in the first step, lignin is decomposed by fungi, and carbon is transferred from the ground litter to the fungal community. Based on the Monod equation, a differential equation for the decomposition of the substrate is established [5] . In this process, the falling rate of herbaceous plants is considered to be a constant, in order to quantitatively calculate the rate of increase of ground litter; in the second step, the fungal community uses the absorbed carbon to multiply and proliferate, and produce metabolites. On this basis, the growth kinetic equation of the fungus was established. Through the above two steps, the growth kinetic model of a single strain is established. Then, the analytical method was adopted to analyze the stability of the differential equations and solve them numerically. Finally, the stable relationship between litter and colony carbon content was obtained [6] . In addition, the stable carbon content of each strain under different temperatures and water potentials has also been analyzed to illustrate the advantages and disadvantages of each strain under different environments. Furthermore, the carbon content trend of various strains in different regional environments can also be predicted [7] . Finally, the stability analysis of the metabolic rate of fungi is carried out, and the change of the stable point of the equations is discussed when the metabolic rate fluctuates within a certain range.

II. THE GROWTH KINETICS MODEL OF SINGLE SPECIES

The complete carbon cycle process in nature is very complicated. Therefore, in this article, the complete process is decomposed, and only the transfer of carbon from the ground litter to the fungus for metabolism is considered. In this segment, the flow of carbon has gone through two processes: firstly, lignin is decomposed by fungi, and carbon is transferred from the ground litter to the fungal community; then, the fungal community uses the absorbed carbon to multiply and proliferate, and produce metabolites.

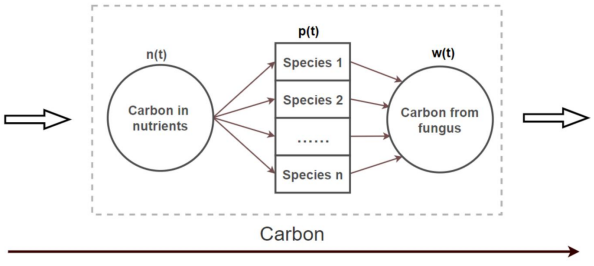


Figure 1. Schematic diagram of carbon flow

Based on the analysis of the mechanism of these two processes, a dynamic equation set consisting of the decomposition equation of the substrate (ground litter) and the improved kinetic equation of cell growth was established. Finally, the stability analysis of the equations and numerical methods are used to study the litter consumption and colony derivation trends under different parameters (temperature, water resistance), so as to study

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00055

the degradation effect of different colonies on the litter under various environmental variables.

A. *Kinetic equation of single species' growth*

The change of total carbon content in the species is composed with two parts: on the one hand, the fungal community constantly absorbs nutrients from the environment and increases; on the other hand, the death of fungi constantly occurs in the community. In the natural environment, it can be assumed that the ground litter is increasing at a constant rate, so the carbon content of litter that can be used by fungi is sufficient. Firstly, a model based on Monod equation is established [5]. In the absence of inhibitors in the medium, the relationship between the specific growth rate of cells and the limiting matrix concentration can be expressed as follows:

$$\mu=\mu_{\max}\frac{n}{n+k}$$

In the above equation, μ stands for the growth rate ratio, $\mu_{\max}$ stands for the maximum specific growth rate, n stands for the limiting substrate concentration, k stands for the saturation constant, of which the value is the limiting substrate concentration when the specific growth rate is half of the maximum specific growth rate.

Based on Monod equation and considering the effects of fungal reproduction and death, a single species growth kinetics equation is established:

$$\frac{dp(t)}{dt}=\lambda\frac{n}{n+k}p(t)-\mu p(t)$$

In the above formula, $p(t)$ means carbon content of the colony; λ means the specific growth rate of the colony; k means the saturation constant and μ means the mortality of the fungal community per unit time.

B. *The decomposition kinetic equation of the ground litter*

As for the ground litter, it can be considered that part of the carbon is absorbed by fungi for proliferation and reproduction, and the other part is decomposed by fungi; at the same time, the ground litter is assumed to be increasing at a constant rate. Therefore, the kinetic equation of litter decomposition was established:

$$\frac{dn(t)}{dt}=-\lambda\frac{n}{n+k}p(t)-Wp(t)+c$$

In the above equation, W represents the metabolic rate of colony, specifically refers to the decomposition rate of substrate nutrients by enzymes produced in colony; c represents the the average rate of herbage withering. Analyzing the above equation, $\frac{dn(t)}{dt}$ stands for the instantaneous change rate of the ground litter carbon content; $\lambda\frac{n}{n+k}p(t)$ stands for the amount of carbon absorbed by the colony in unit time, which will lead to the decrease of litter carbon content; $Wp(t)$ stands for the consumption of litter carbon by colony metabolism, which will further lead to the decrease of litter carbon content. c was used to represent the average rate of herbage withering, which will lead to the increase of the ground litter carbon content.

C. *The decomposition kinetic equation of the ground litter*

Combined with the above analysis, a complete set of single species' growth kinetics equations is obtained as follows:

$$\begin{cases}\dfrac{dn(t)}{dt}=-\lambda\dfrac{n}{n+k}p(t)-Wp(t)+c\\[2ex]\dfrac{dp(t)}{dt}=\lambda\dfrac{n}{n+k}p(t)-\mu p(t)\end{cases}$$

In the above equations, the first equation is used to measure the change of carbon content of fungal community, and the second equation is used to measure the change of carbon content of ground litter. By calculating the stable point of the second-order differential equations, the growth state and litter decomposition efficiency of a single fungal community in the natural environment can be studied.

D. *Stability analysis of growth kinetics equations*

To study whether the dynamic system is stable through the Kinetic equations, stability analysis is very crucial. For a system of second order differential equations, when $\frac{df(n^*,p^*)}{dn}+\frac{dg(n^*,p^*)}{dp}<0$ and $|J|>0$, the corresponding equilibrium point can be obtained as (n^*,p^*), which is stable. Through calculation, it can be concluded that:

$$\begin{cases}\dfrac{df(n^*,p^*)}{dn}+\dfrac{dg(n^*,p^*)}{dp}=-\dfrac{(\lambda-\mu)^2c}{\lambda k(w+\mu)}<0\\[2ex]|J|=\dfrac{(\mu+w)(\lambda-\mu)^2c}{\lambda k(\mu+w)}>0\end{cases}$$

Based on the above results, the equilibrium point can be considered stable

III. THE SETTLEMENT OF THE MODEL

The data on the relationship between the growth rate of different strains and the ambient temperature are found in Literature [7], which is as follows:

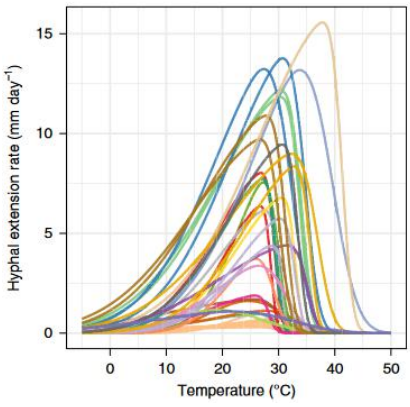


Figure 2. Curve of mycelial elongation with temperature

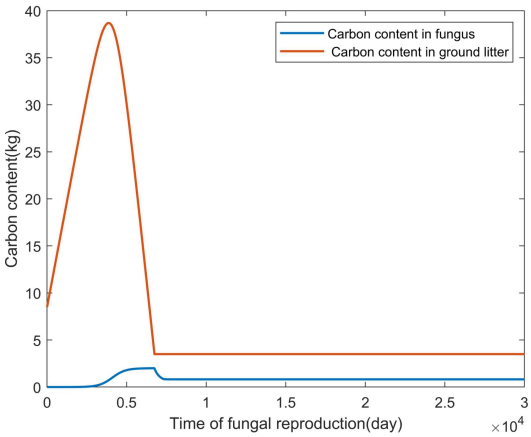
From the above images, with the increase of temperature and moisture, the growth rate of most species of fungi increase at first and then decrease. The reasons can be summarized as follows: before the temperature T reaches the optimal growth temperature T_0, the growth rate λ increases with the increase of T, and reaches the peak at $T = T_0$, and then decreases with the increase of T until T reaches the maximum temperature $T_{\max}, \lambda = 0$, when the fungi is dead. Besides, due to different degrees of environmental adaptation to different fungi species, the suitable temperature ranges for fungi species also show great differences.

On the basis of establishing and analyzing the above equations, the species **a.gal1.s** is chosen as an example. Fixing the saturation constant k, fungal metabolic rate w, the mortality rate of fungi μ, the temperature T, the water potential m, the dropping rate c, the initial conditions p_0, n_0 on a specific and reasonable value, as are shown in the table below:

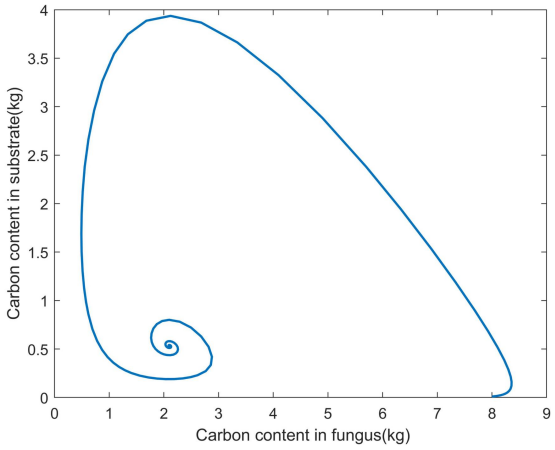
TABLE I. VALUES OF PARAMETERS AND CORRESPONDING UNITS

Parameter	Values	Unit
k	10^{-3}	—
w	0.01	—
μ	0.01	—
T	5.87	℃
m	-0.43	*MPa*
c	0.04	*kg / day*
p_0	0.001	*kg*
n_0	5	*kg*

According to the cited data, it is available that the growth rate of this species under this initial condition is: λ = 0.02. To settle out the model,numerical solution was used and Growth curves of fungi as well as substrate (ground litter) consumption curves (5000 days reproduction of fungi) was drafted as follows:



(a) Trends of litter and fungi



(b) phase diagram

Figure 3. Results of the model

It is known from the figure that when the fungal reproduction number was 7320, the carbon content p of the colonies gradually leveled off, implying that the fungal growth and reproduction reached an equilibrium point at 7320 days, when the fungal birth rate and mortality rate were leveled off. At this time, the decreasing trend of the carbon content in the litter also tended to level off. As a result, the conclusion was obtained: the balance point of fungal growth and the balance point of litter consumption are the same with each other, which is also in accordance with the actual situation.

A. The impacts of different temperatures on the same species

When keeping other variables and parameters in the environment unchanged, and only changing the ambient temperature, the time for the same strain to reach a stable point is very different. In order to explore the specific influence of temperature on the decomposition rate of fungi, a representative fungus **a.gal1.s** was selected as the research object. 5 sets of data on the relationship between the number of generations required for the fungus to reach a stable point and the temperature are sorted out, importing the data into MATLAB and drawing the image. The result is shown in Figure 4.

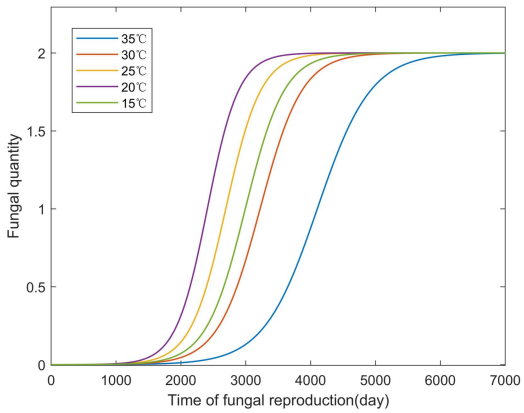
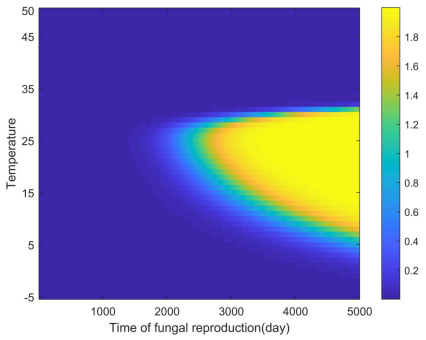


Figure 4. The relationship between fungal growth and temperature

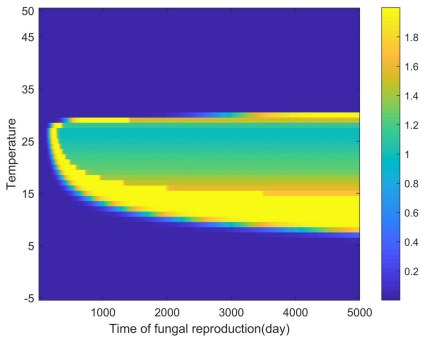
Because the overall activity and growth rate of the fungus is better between 15℃ and 35℃, temperature in this interval are chosen to study the changes in fungal carbon content. Analyzing the above figure, it can be found that under different ambient temperatures T, the time for fungus **a.gal1.s** to grow and reach the environmental capacity is different, and the required time decreases first and then increases as the ambient temperature T increases. After reaching a stable point for a period of time, the nutrients are consumed, and the fungal weight drops rapidly.

B. *The influence of temperature changes on different species*

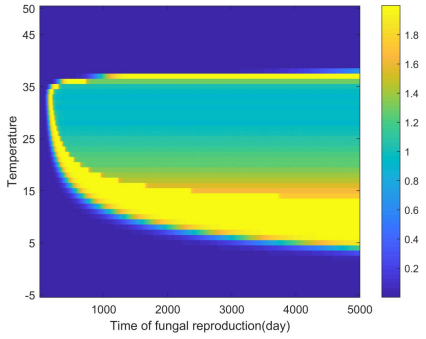
To explore the effects of temperature changes on different fungi, 3 representative species (fungus **a.gal1.s**, fungus **I.carib.s**, fungus **s.comm.n**) were selected for analysis and the heat map is drafted to show the time required for the growth of each species to reach equilibrium at different temperatures, as shown below:



(a) Heat map of fungus **a.gal1.s**



(b) Heat map of fungus **I.carib.s**



(c)Heat map of fungus **s.comm.n**

Figure 5. The relationship between fungal growth and temperature

The dark blue area indicates the period when the overall colony mass is small and the yellow area indicates the state after the colony has grown to a stable point. After the nutrients are gradually consumed, the total colony mass drops rapidly from the stable point. Observing the stable points of different strains at different temperatures, the most suitable temperature corresponding to the fastest time for each strain to reach the stable point can be found. Therefore, the optimum temperature for different strains is different

C. *Summary*

After analyzing the temperature changes on the fungus **a.gal1.s**, fungus **I.carib.s** and fungus **s.comm.n**, the following conclusions are obtained: The optimum temperature for different fungus to reach a stable point is completely different, and there are differences in the advantages and disadvantages of the growth of each fungus under different temperature conditions.

IV. CONCLUDING REMARKS

Based on the carbon cycle, this paper discusses the fungal decomposition process of ground litter and establishes the corresponding differential equation model. The result of stability analysis proves that this model has excellent properties. Furthermore, this article discusses the influence of external environment temperature on the growth of fungi and the complex relationship between them, which is innovative and practical.

V. ACKNOWLEDGEMENT

I am sincerely thankful to professor Shesheng Zhang for some valuable suggestions in the process of writing the paper, which have greatly improved the quality of the paper.

REFERENCES

[1] Yuning Li, Guanyu Wang,Wei Li.Soil Respiration and the Global Carbon Cycle[J]. Earth Science Frontiers(in Chinese), 2002(02): 351-357.

[2] Gude,S.,Pinçe,E.,Taute,K.M.,Seinen,A.(2020).Bacterial coexistence driven by motility and spatial competition.Nature, 578(7796),588-592

[3] M. Lustenhouwer,N.,Maynard,D.S.,Bradford,M.A.(2020).A trait-based understanding of wood decomposition by fungi.Proceedings of the National Academy of Sciences

[4] Wei Hong,Chengzhen Wu,Shujun Yan.Modification of Population Growth Model[J],Chinese Journal of Applied and Environment Biology(in Chinese),2004(01):23-26.

[5] Li,R.G.,Zhang,Y.Q.(2017).A BOD Kinetic Model for Municipal Sewage Based on Monod Equation.

[6] Yuan Tian,Keyou Wang,Jin Xu,Guojie Li.Construction of Spatial Transient Stability Boundary for Power System Parameters and Online Quick Update Method[J].Automation of Electric Power Systems(in Chinese):1-15.

[7] Maynard, D.S., Bradford,M.A., Kristofer,C.R.(2019). Consistent trade-offs in fungal trait expression across broad spatial scales.Nature Microbiology,04(5),846-853.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Research on multi-level malicious Web page Recognition based on topic Fusion and improvement of CNN

Zhou Huang
School of Computer Science and Technology,
Wuhan University of Technology
Wuhan, China
e-mail: 1239746898@qq.com

HanBing Yao
Wuhan University of Technology
Chongqing Research Institute
School of Computer Science and Technology,
Wuhan University of Technology
Wuhan, China
e-mail: 22876681@qq.com

***Abstract*—As one of the portals of the Internet, web application brings convenience to human society, but also has many problems,such as phishing attacks, malware downloads, privacy breaches, etc.In order to more accurately identify malicious web pages, reduce the risk of network attacks,this paper designs a multi-level detection model according to the different level characteristics of web pages.The model uses LDA-SECNN to learn the characteristics of web page content, uses random forest algorithm to learn the static characteristics of web page, and finally combines the two outputs to determine.Experiments show that the multi-level model can improve the accuracy of the model, and has good stability and convergence.**

***Index Terms*—Malicious web pages; LDA theme; Convolutional neural network; Random forest**

I. Introduction

blahblah Malicious web pages refer to a collection of web pages that appear in the form of web pages, steal user privacy, install malicious programs or execute malicious code during access.In recent years, with the growth of the types and number of web pages, the frequency of malicious web pages has also increased greatly.At the same time, due to the improvement of the attacker's technical level, the content of malicious web pages is more confusing, the malicious behavior of web pages is more hidden, and it is easy to induce users to carry out unsafe operations.Malicious web page detection is a very important part of network security.

According to different malicious web page behaviors, corresponding solutions are proposed.Literature [1] uses complex recursive units to train two recursive neural networks to detect abnormal network requests. Compared with the most advanced model, it has achieved better results and can reduce the impact of feature selection.Literature [2] fully considers the relationship between word position and context in URL, uses special characters to segment URL, and then uses convolutional neural network as classifier. The experimental results show that the model has achieved good results on a large number of real data sets.Literature [3] designed a comprehensive web page association analysis model by using the technologies of topic tracking, topic anomaly discovery, web page visual similarity evaluation, web page structure analysis and URL analysis, which solved the problem of difficult unknown feature detection.Reference [4] proposed a multi-scale feature fusion detection method, which models HTTP requests from word level and character level. With the help of multi-scale feature fusion technology, it has achieved good results on real data sets.Literature [5] designed a multi-level classifier for different levels of web page features, and combined with the output of neural network and random forest for judgment. Experiments show that the multi-level model can improve the recognition rate of the model.

Through the above research, it is found that due to the complex structure of web pages, it is difficult to take into account the characteristics of different levels of web pages. The general model usually models web pages'URLs, themes, static features, and so on,Through the study of literature [5], we can see that model fusion can make up for the defects of single model and improve the overall performance of the model.Therefore, a multi-level recognition model based on topic fusion and improved convolutional neural network is proposed in this paper. Firstly, the word2vec word vector corresponding to each word of web page content is fused with LDA topic vector, and a word vector fusion algorithm is proposed according to the contribution of words to web page topic category; Then, senet is introduced into CNN to build a web page topic recognition model secnn, which can improve the performance of important features and suppress useless features, and enhance the ability of feature extraction; At the same time, the random forest classification model is used to learn the static characteristics of web pages, and finally the two outputs are synthesized for judgment.

II. Model Design

A. Word vector fusion algorithm

There are some problems in word2vec vector representation, such as the lack of global semantic expression and the inability

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00056

to highlight the importance of words in document categories. This vector indicates that there are problems such as the lack of global semantic expression and the inability to highlight the importance of words in the document category.The topic vector can complete the global semantic expression through the probability distribution of the words in each topic. The integration of Word2vec and LDA can better represent the text. This article uses the Word2vec fusion LDA method to improve the web topic recognition effect.First extract the URL part, content part, and JavaScript part of the web page to form an input document, then normalize the Word2vec word vector and LDA body vector corresponding to each word, and then the vector fusion vector is based on the contribution of the word to the category.The degree is weighted to highlight the importance of words in the document category.The vector fusion algorithm proposed in this paper is as follows:

(1)Use the CBOW model in Word2vec to train the corpus to obtain the word vector, d represents the dimension of the word vector, and $|P|$ is the vocabulary of the entire corpus trained with the word vector.

For the word $word_j$, its corresponding Word2vec word vector is expressed as $\vec{x}_j = (p_{j,1}, p_{j,2}, \cdots, p_{j,d})$.

(2)The topic-word matrix φ obtained by using the Gibbs sampling method to train the LDA topic model on the corpus is shown in equation (1).

$$\varphi = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,|\varphi|} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,|\varphi|} \\ \cdots & \cdots & \cdots & \cdots \\ z_{o,1} & z_{o,2} & \cdots & z_{o,|\varphi|} \end{bmatrix} (1)$$

$|\varphi|$ represents the size of the vocabulary of the topic model, and O represents the dimension of the vector, which is the same as the dimension setting of the word vector. For the word $word_j$, the corresponding topic vector is denoted as $\vec{y}_j = (z_{1,j}, z_{2,j}, \cdots, z_{o,j})$.

(3)Reference [6] mentions the explanation of word vectors, that is, word vectors can express semantics, and the addition of word vectors can also express the combination of word meanings. Therefore, this paper adopts the vector fusion method of addition to normalize the Word2vec word vector and topic vector corresponding to each word and add them to form a new vector representation.$\|\cdot\|_2$ represents the 2-norm, and the new vector representation of the word $word_j$ is shown in formula (2).

$$v'_j = [\frac{\vec{x}_j}{\|\vec{x}_j\|} + \frac{\vec{y}_j}{\|\vec{y}_j\|}](2)$$

A document is represented by the word vector corresponding to formula (3) as $d_i = [v'_1, v'_2, \cdots, v'_m]^T$, and m is the sentence length of the document.

(4)Use the graph-based TextRank algorithm to extract keywords from document d_i, map each word in the document as a graph node, and map the co-occurrence relationship of the words in a preset window to the edges of the nodes,initialize the initial weightsw_{ji} of V_j nodes and randomly,According to formula (3), iteratively calculate the weight of each node until convergence, $a$represents the damping coefficient, and $nt(V)$ represents the set of adjacent nodes of node V.

$$s(V_j) = (1-a) + a \times \sum_{V_i \in nt(V_j)} \frac{w_{ji}}{\sum\limits_{V_t \in nt(V_i)} w_{it}} s(V_i)(3)$$

After obtaining the weight of the words in the document, in order to simplify the calculation, only the top k words in the weight ranking are taken out to form the keyword set $kword = \{kd_1, kd_2, \cdots, kd_k\}$.For each document, the weight factor $g(word_j)$ of $word_j$ is obtained by formula (4), and the weight factor represents the importance of the word.

$$g(word_j) = \begin{cases} s(word_j) \times \eta_j & word_j \in kd \\ 1 & word_j \notin kd \end{cases} (4)$$

Calculate the topic weight corresponding to the keywords in the text. The topic weight reflects the distribution of words in different topics. The larger the topic weight, the greater the difference in the distribution of words in each topic, and the corresponding contribution to topic recognition The greater the degree, the calculation of the topic weight is shown in equation (5).

$$\eta_j = \frac{\sqrt{\sum\limits_{i=1}^{b} \left[df(b_i, word_j) - \frac{df(D, word_j)}{b} \right]^2}}{b} (5)$$

Among them, η_j represents the topic weight of the word $word_j$, b represents the number of topic categories, and $df(b_i, word_j)$ represents the number of texts containing the word $word_j$ in category b_i,$df(D, word_j)$represents the number of texts containing the word $word_j$ in the entire corpus D.

The weighted word vector of the word $word_j$ is expressed as shown in equation (6).

$$v_j = g(word_j) \otimes v'_j (6)$$

Among them, v_j represents the weighted word vector, the symbol $\otimes$ represents element-wise multiplication, and the weighted word vector of the document d_i is expressed as: $d_i = [v_1, v_2, \cdots, v_m]$.

The document vector obtained by this algorithm can represent the text better than the general Word2ve word vector representation. Using the document vector obtained by this algorithm as the feature representation of the neural network for learning is helpful to improve the performance of the model.

B. LDA-SECNN Web Page Topic Recognition Model

With the more and more extensive application of deep learning in NLP field[7][8], recurrent neural network (RNN) and CNN model have been widely used in the field of text classification. According to the research of Yin et al[9]., the training of CNN model has more advantages in text classification, so we decided to adopt CNN model.

After vector fusion, the semantic feature distribution of the document is unstable. In order to make the model pay more

attention to the features beneficial to the topic recognition task, the se module in senet is introduced into the convolution neural network, and a secnn model suitable for the topic recognition task is designed. The structure diagram of secnn model is shown in Figure 1.

The number of channels for each height of the convolution kernel output feature map is C, and the SE layer uses the feature map output by the convolution layer as output. First, perform a global average pooling operation on each feature channel Y_C of the feature map to obtain the compressed The feature map Z_C is shown in formula (7).

$$z_c = F_{sq}(Y_c) = \frac{1}{W' \times H'} \sum_{i=1}^{W'} \sum_{j=1}^{H'} Y_c(i,j) (7)$$

Where W',H' represent the size of the feature map,$c = \{1, C\}$.

Then through two fully connected operations, training parameters W_1 and W_2 and the ability to recognize the importance of each feature channel, highlight the important features for the topic recognition task,The normalized weight vector S is obtained by the Sigmoid function as shown in equation (8).

$$S = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) (8)$$

Among them, δ represents the ReLU activation function, σ represents the Sigmoid activation function,$W_1 \in R^{\frac{C}{r}C}$, $W_2 \in R^{\frac{C}{r}C}$,r represents the number of hidden layer nodes in the middle layer.

Finally, the weight vector is weighted to the original feature map channel by channel through multiplication, and the final output Y_c is obtained. The original feature is recalibrated in the channel dimension. The weighted feature map and the original feature map have the same dimension, as in formula (9) Shown.

$$Y_c = F_{scale}(Y_c, S_c) = S_c \cdot Y_c (9)$$

Among them,S_c represents the weight vector corresponding to the response feature map, and the symbol $\cdot$ represents the vector product operation.

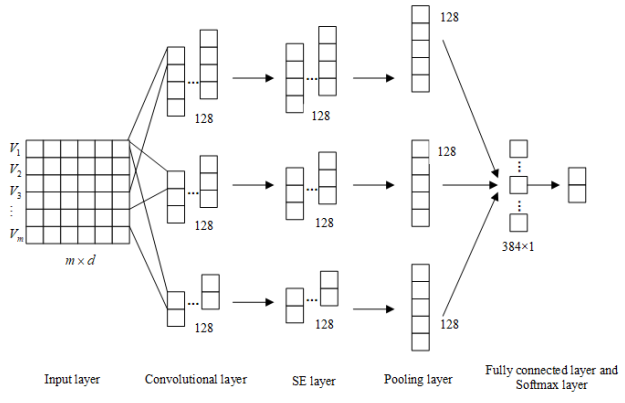


Fig. 1. LDA-SECNN structure

TABLE I

Page features	
Total URL length	URL path part length
Whether to use https protocol	Are there uncommon ports
Does it contain '@'	The number of characters '.'
Number of keywords	Number of top-level domains
Whether to use iframe	Domain name survival time
URL path part length ratio	Whether to include ip address
URL information entropy	The number of characters '-'
DNS records	

C. Webpage Static Feature Recognition Model

The random forest algorithm is an integrated machine learning method that can be used to control overfitting. In addition, with its forest structure, the instability of individual decision trees may disappear. Therefore, this experiment uses the random forest algorithm to detect web page features. The classifier first extracts web page features and generates a training set. Then randomly select a number of training subsets, and construct a decision tree for the training subsets. Finally, the output of the decision tree is obtained, and the average value of the output is obtained to obtain the detection result of the webpage.

As shown in Table I, the article extracts 15 features of web pages, including URL string features, DNS information, and web HTML features.

D. Threshold setting

As shown in Figure 1, the threshold α determines whether the web page performs the second level of detection.As shown in the formula, if the ratio of the maximum and minimum values between the normal webpage probability p1 and the malicious webpage probability p2 output by LDA-SECNN is less than α , then web URL features, web HTML features, and DNS information need to be extracted for the second level of detection;otherwise, the output result of LDA-SECNN will be directly judged. α is initialized to 1, and then the recognition accuracy rate of the multi-level detection model is output, α is added by 1, and then the optimal recognition accuracy rate is output and converged.

$$\begin{cases} \frac{\max(p_1,p_2)}{\min(p_1,p_2)} > \alpha, Direct judgment \\ \frac{\max(p_1,p_2)}{\min(p_1,p_2)} \leq \alpha, \text{Sec}ondary\text{detection} \end{cases}$$

III. Experiments and Analysis

A. Evaluation index

This article uses accuracy (A), precision (P), recall (R) and F1 value to evaluate the performance of the model. Among them, accuracy is the most important evaluation index. The precision rate indicates the proportion of the samples that are predicted to be positive, and the recall rate indicates the proportion of the positive samples that are predicted to be correct. The F1 value is the harmonic average of the recall rate and the precision rate.

TABLE II

algorithm	A	P	R	F1
CNN	95.68	95.53	95.78	95.66
LSTM	95.21	95.90	94.43	95.16
LDA-SECNN	97.22	96.94	97.45	97.20

$$A = \frac{TP+TN}{TP+TN+FN} \times 100\%$$
$$P = \frac{TP}{TP+FR} \times 100\%$$
$$R = \frac{TP}{TP+FN} \times 100\%$$
$$F1 = \frac{2\times P\times R}{P+R} \times 100\%$$

In the above formula, TP represents the total number of positive samples predicted as positive samples, FP represents the total number of negative samples predicted as positive samples, and TN represents the total number of negative samples predicted as negative samples. , FN indicates that the negative samples are predicted, which is actually the total number of positive samples.

B. Experimental environment and data set source

The programming language used in this experiment is Python3.5, the CPU is AMD Ryzen 7 4800U with Radeon Graphics 1.80 GHz, and the RAM is 16G. Part of the data of benign webpages comes from websites published by Alexa. In order to ensure the purity of benign webpages, the top 1000 websites are selected. At the same time, 20,634 malicious web pages were crawled from Phishtank, as well as a web page data set from the data science competition platform kaggle[10]. After deduplication and cleaning, the final web page data set contained 36,696 benign web pages and 35,478 malicious web pages. For LDA-SECNN, set the batch bit to 64, epoch to 20, and use 5-fold cross-validation during the experiment to ensure the stability of the model.

C. Result analysis

As shown in Table II, compared with CNN and LSTM detection models, LDA-SECNN is better than the former in accuracy, recall, precision, and F1 value. Through the comparison of Figure 2, the multi-level detection model designed in this paper performs better on indicators such as accuracy. As shown in Figure 3, when the value of α is 219, the recognition accuracy of the multi-level detection model converges, and it is significantly better than the LDA-SECNN model and the RF model.

IV. Conclusion

This paper integrates the content of the webpage into the word vector expression, designs a multi-level detection model combining neural network and random forest algorithm, and compares it with the single-level detection model. Experiments have proved that the multi-level detection model has better results in terms of accuracy, F1 value and other indicators. There are many types of malicious webpages. In the future, we can study the multi-classification of malicious webpages, which is helpful for targeted defense.

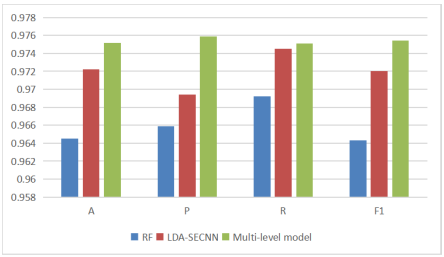


Fig. 2. Comparison chart of RF, LDA-SECNN, multi-level detection model

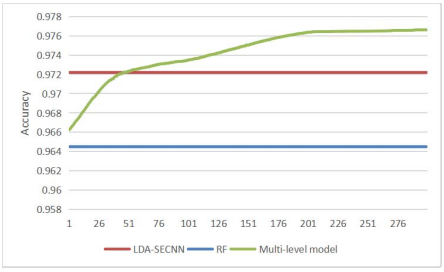


Fig. 3. Threshold α change graph

Acknowledgments

The work described in this paper was supported by the Research Project of Wuhan University of Technology Chongqing Research Institute (YF2021-10).

References

[1] Jingxi Liang, Wen Zhao, and Wei Ye. 2017. Anomaly-Based Web Attack Detection: A Deep Learning Approach. In Proceedings of the 2017 VI International Conference on Network, Communication and Computing (ICNCC 2017). Association for Computing Machinery, New York, NY, USA, 80–85. DOI:https://doi.org/10.1145/3171592.3171594

[2] Wu Haibin,Zhang Dongmei. Malicious URL detection technology based on context information[J]. Computer engineering Software,2019,40(1):63-68. DOI:10.3969/j.issn.1003-6970.2019.01.013.

[3] Senhao Wen, Zhiyuan Zhao, and Hanbing Yan. 2018. Detecting Malicious Websites in Depth through Analyzing Topics and Web-pages. In Proceedings of the 2nd International Conference on Cryptography, Security and Privacy (ICCSP 2018). Association for Computing Machinery, New York, NY, USA, 128–133. DOI:https://doi.org/10.1145/3199478.3199500

[4] WuJiahong,Zhen Guo Yang,Liu Wenyin.Malicious HTTP request detection method based on multi-scale feature fusion[J].Application Research Of Computers,2021,38:871-874+880.

[5] Zhang Shikun.Research on malicious web page detection technology based on multi-layer classifier[J].Modern computer,2020(18):64-68.

[6] ZHOU Wanting. Research on vector representation of text sentiment analysis based on word vectors[D]. Changchun: Northeast Normal University, 2019.

[7] Johnson R, Zhang T. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks[J]. Eprint Arxiv, 2014.

[8] Wen Y, Zhang W, Luo R, et al. Learning text representation using recurrent convolutional neural network with highway layers[J]. 2016.

[9] Yin W, Kann K, Yu M, et al. Comparative Study of CNN and RNN for Natural Language Processing[J]. 2017.

[10] Singh A. K. (2020). Malicious and Benign Webpages Dataset. Data in brief, 32, 106304.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Application and Research of Intelligent Security System Based on NFC and Cloud Computing Technology

Hua Jiang
Department of Information Engineering
Shandong Vocational College of Science and Technology
Weifang City, Shandong Province, China
wf-jianghua@163.com

***Abstract*—With the rapid development of urbanization, community security and public security have become an important social issue. As conventional patrol methods can not effectively ensure effective supervision, this paper studies the application of NFC (Near Field Communication) technology in intelligent security system, designs and constructs a set of intelligent security system suitable for public security patrol or security patrol combined with current cloud service technology. The system can not only solve the digital problem of patrol supervision in the current public security, but also greatly improve the efficiency of security and improve the service quality of the industry through the application of intelligent technology.**

Keywords-component; NFC; Cloud computing technology; intelligent security system

I. BACKGROUND

Community security patrol is an important social security work. Both police and security personnel need to conduct periodic patrol security work in a certain area. With the rapid development of urbanization, the community is becoming more and more dense. At the same time, the uncertainty of floating population and the frequent occurrence of theft and robbery in the community make people pay more and more attention to the security and public security of the community.

The conventional patrol mode is nothing more than through the movement of personnel. Whether the patrol personnel cheat or hide favoritism, etc. can not be effectively supervised.

NFC technology is a short-range high-frequency wireless communication technology. At present, most smart phones in the market have NFC technology. Through NFC technology, short-distance communication between two settings can be realized, which is very suitable for security work.

Through NFC technology, patrol personnel can hold mobile phones where they need to sign in. Then, check in with the NFC function of the mobile phone. This data will be displayed in the background, which point the patrolman patrols to and when he patrols, including his behavior track. If the public security patrols and the alarm receiving center receives the alarm, it can quickly know which police are patrolling at the alarm point, and the police force can be deployed nearby immediately. NFC is real-time check-in, which can greatly improve the authenticity of patrol and make the emergency allocation of emergencies more efficient. In this way, in the construction of smart city, the duty support of public security or security personnel is greatly improved through the smart patrol system of NFC technology.

II. TECHNICAL INTRODUCTION

A. NFC Technology

Near field communication (NFC) has evolved from non-contact radio frequency identification (RFID), which is based on RFID and interconnection technology. Near field communication is a short-range and high-frequency radio technology. Like RFID, the near-field communication information is also transmitted through the electromagnetic induction coupling mode of the wireless frequency part of the spectrum, but there is still a big difference between the two. The transmission range of near-field communication is smaller than that of RFID, and the transmission range of RFID can reach 0-1m. However, due to the unique signal attenuation technology adopted in near-field communication, compared with RFID, near-field communication has the characteristics of low cost, high bandwidth and low energy consumption. Therefore, NFC technology has been supported by more and more manufacturers, and the chip price will be cheaper and more widely used. At present, NFC technology is mainly used in the mobile payment function, and the patrol field in the security industry is still in the initial stage of development.

In the intelligent security system, patrol personnel exchange, store and display real-time data with the background through the smart phone with NFC chip, which greatly improves the authenticity of patrol.

B. Cloud computing technology

Cloud computing is distributed computing, parallel computing, utility computing, network storage technologies, virtualization, load balance, and high Available) and other products of the integration of the development of traditional computer and network technology. It is a supercomputing mode based on the Internet. Cloud computing can be considered to include the following layers of services: infrastructure as a service (IAAs), platform as a service (PAAS) and software as a service (SaaS).

In the intelligent security system, the NFC patrol data and location data of patrol personnel are used as data sources to analyze the data through the cloud computing platform.

C. Artificial Intelligence Technology

Artificial intelligence (AI). It is a new technology science to research and develop the theory, method, technology and application system for simulating, extending and expanding human intelligence. The research in the field of artificial intelligence includes robot, language recognition, image recognition, natural language processing and expert system. Since the birth of artificial

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00057

intelligence, the theory and technology are increasingly mature, and the application field is also expanding.

In the intelligent security system, the location data and time data of patrol personnel are used as data sources. Using artificial intelligence method to analyze data, optimize patrol path and patrol time, and improve security efficiency.

III. DESIGN OF SECURITY INSPECTION SYSTEM

A. System framework

Starting from the existing patrol system, cloud computing technology, NFC technology, artificial intelligence and other aspects, this paper proposes a security patrol system architecture suitable for community security, and verifies the security and effectiveness of the architecture in part, providing reference for other researchers, and providing basis for further research.

The system framework is shown in Figure 1.

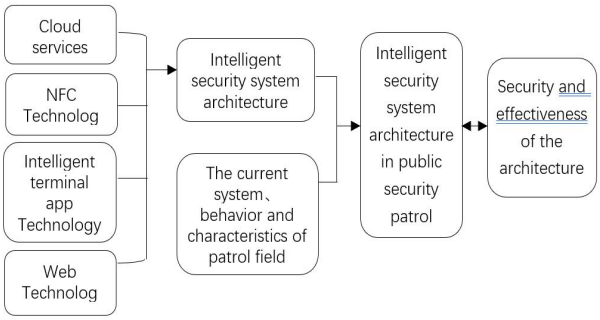


Figure 1. The system framework

B. Patrol data acquisition system based on NFC

The data collection of Community Patrol Inspection consists of mobile phone with NFC reader and tag card. NFC can only communicate with tag card in close range. For example, the value of this project is set to be less than 10cm, which basically eliminates the on-site cheating of patrol personnel. After the inspectors log in to the system, click the send data button, and the mobile phone will send the basic information and location information to the cloud computing center by 4G communication signal colleagues. The flow chart of NFC data acquisition is shown in Figure 2.

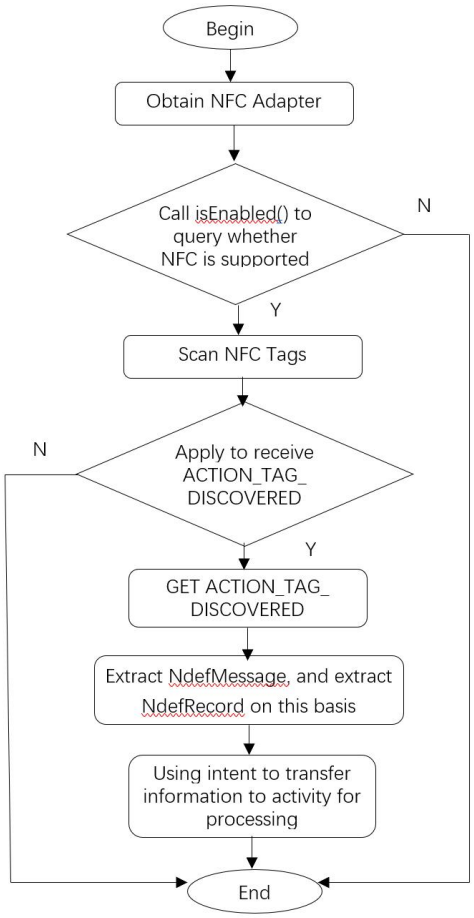


Figure 2. The flow chart of NFC data acquisition

C. Cloud Service Platform

In order to mine and analyze patrol data, based on Hadoop open-source cloud computing platform, cluster analysis, association analysis and in-depth learning methods are built under MapReduce architecture. The location information, path information, personnel information and other data in the patrol process are analyzed and processed to obtain the optimal configuration of personnel and path in the patrol process, and the query level of patrol information is built . There are four layers in this system platform. They are human-machine interface layer, data Processing layer, database layer and data integration layer. The system platform architecture is shown in Figure 3.

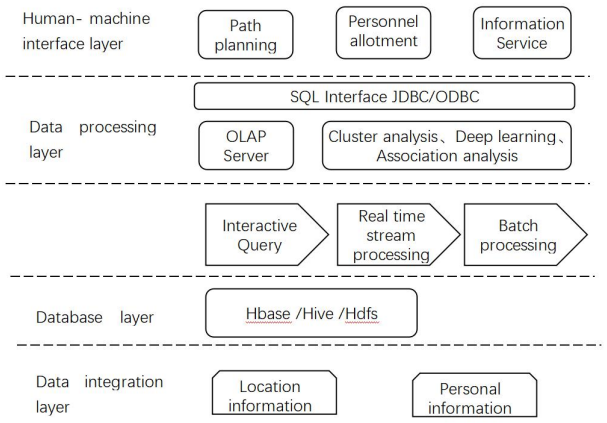


Figure 3. The system platform architecture

IV. SUMMARY

This paper analyzes NFC technology, cloud service technology and terminal app related technology, combined with the characteristics of public security patrol, designs a smart security patrol system. The system can not only solve the problem of digital patrol supervision in the current public security, but also greatly improve the efficiency of security and the service quality of the industry through the application of intelligent technology.

The research results of this topic can not only be used as a supplement to the current public security patrol, improve the security and effectiveness of security, but also as one of the standards of the next generation of patrol system, which has great use and practical value.

[1] Lin Qing. Smart application of big data security based on cloud computing [J]. China Public Security (Comprehensive Edition), 2013, 000 (020): 188-189.

[2] Wang Jingbin. Upgrading and application of big data security in smart city construction [J]. China Public Security (Comprehensive Edition), 2018 (08): 58-60.

[3] Chen Zhenghui. Research on security equipment inspection system based on NFC technology [J]. Mobile information, 2020 (1): 00057-00059.

[4] Chen Zhipeng, Wang Bin. Design of intelligent patrol inspection system based on NFC technology [J]. Internet of things technology, 2017, 7 (002): 24-26.

[5] Chen Zhipeng, Wang Bin. Design of intelligent patrol system based on NFC technology [J]. Internet of things technology, 2017, 7 (002): 24-26. [1] Zhang Ming, Zheng Ziwei. Research on mobile patrol technology based on NFC [J]. Data communication, 2018, 000 (001): 8-11,21

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Research on Network User Behavior Management System Based on Blockchain Technology

Sun Jianjun

Admissions and Employment Office
Shandong Vocational College of Science and Technology
Weifang, China
e-mail:631901036@163.com

***Abstract*—In the information age, the network environment has become the main platform for information exchange and transmission. The complexity of Internet information and the diversification of network user behavior have brought many difficulties to the Internet information management. Standardization of network user behavior and analysis and management of network information have become an important research hotspot in the Internet age Based on the characteristics of distributed storage, this paper proposes a network user behavior management system based on blockchain technology, which can improve the information management and behavior constraints of the network environment on the premise of ensuring the free behavior of network users.**

Keywords-Network user behavior management; Blockchain; Smart contractn

I. INTRODUCTION

With the development of the Internet era, the network user behavior is increasing rapidly, and its information release, information interaction, information transmission have become the main user behavior in the network environment. Due to the large space of the network platform, the freedom of user behavior is relatively high, and user behavior is prone to violation of network security, laws, regulations, humanitarian and other related constraints.The research and management of user behavior in the network environment is the main subject after the development of Internet. According to the activity information and related trajectory of users in the network environment, the user behavior in the network environment can be analyzed and managed. As the mainstream of current new technology development, blockchain technology has the advantages of decentralization, tamper proof, distributed and high data security. It has been integrated with finance, economy, education and other industries, providing new ideas for data analysis and management. In this paper, according to the characteristics of network user behavior management, combined with the advantages of blockchain technology, the application of blockchain distributed data storage, data encryption technology, intelligent contract technology and Internet user network behavior management, constrains network user behavior, analyzes user behavior rules, filters data information, and proposes an innovative network user behavior management system Information management provides technical reference[1].

II. INTRODUCTION TO THE BLOCK CHAIN AND RELATED TECHNOLOGY

The block chain technology originated from the development of Bitcoin, mainly working with the bottom of Bitcoin technology.Block chain technology can be defined as: distributing and updating data using a point-to-point distributed technology, using the encrypted chain block structure to store and verify data, with a detrimentary architecture of the intelligent contract as a means of managing data[2]. And the block chain is a new infrastructure and distributed calculation paradigm.As a technical solution, block chain technology integrates a variety of types of technologies and innovative technology models.

The core of block chain technology has mainly distributed book technology, asymmetric encryption algorithm, intelligent contract technology. The distributed book technology mainly combines the block chain as a book, and the node of the entire distributed environment completes the bill of transaction, and the accounts recorded by each distributed node are complete, and the entire distributed environment is used. The centralized method, breaking the traditional administrator's one-point accounting method, reflecting the integrity of its transactions; the asymmetric encryption algorithm is encrypted, decrypting the data using the public key and the private key, thus improve data transmission, storage security; intelligent contracts exist in the form of computer programs, with an embedded programming contract, mainly running in a shared area block chain data accounting environment, contract participants together, code scripts formed It is a transparent form storage. After debugging, the deployment is deployed on the block chain, improves the credibility of the program, and data operation[3].

Decentralization, consensus mechanism, traceability, and openness are the main characteristics of block chain technology. Decenturing refers to the permissions to distribute records and storage on all systems in the block chain. Each node in the block chain communicates with each other, presenting a mesh information structure, implementing verification, transmission, and data management of data. There is no one, and the permissions and tasks of all nodes are equivalent to the task, and the

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00058

entire functional implementation of the joint maintenance system is implemented. The consensus mechanism refers to the characteristics of the district block chain technology through intelligent contract, and improve reliability and credit by means of corresponding technical rules. Traceability technology refers to the time dimension of the block chain to embed the time dimension in the block chain to record the specific time of the transaction[4]. Finally form a time chain, making the data high traceability.

As one of the block chain technologies, the consensus mechanism refers to the process of reaching a unified protocol on the status of the network in the way, and is also known as consensus algorithms, which helps verify and verify information to be added to the classified book. Make sure that only real transaction is recorded on the block chain. The consensus mechanism has the characteristics of the block chain to centralize. Since the consensus mechanism is mainly for the purpose of unified consensus in the system, the consensus mechanism has reached a unified protocol for the authenticity and accuracy of the data. When the system data is updated, the consensus mechanism is responsible for securely updating the data status in a distributed network. The protocol rules embedded in the network ensure that the state of the public classification account is always updated with the public's consensus.

III . FRAMEWORK OF NETWORK USER BEHAVIOR MANAGEMENT SYSTEM BASED ON BLOCKCHAIN TECHNOLIGY

In this paper, combined with blockchain data encryption, distributed storage, consensus mechanism and smart contract sub technologies, the network user behavior management system based on blockchain technology is proposed. The system can intelligently screen, mark and manage user behaviors such as information release, storage and data sharing after network users log in, so as to realize the management and analysis of user behavior in the network environment on the basis of purifying the internet environment, improve the intelligence and security of user management[5].

Figure 1 shows the network user behavior management system architecture under block chain technology, and the entire system is divided into application layer, protocol layer, data encryption layer, and distributed storage four layers. The application layer provides user registration and information registration, with the browser or client to display the system entry and present user behavior information. After the network user is registered by the unique username identifier, the behavior trajectory in its network environment is recorded. The user's behavior instruction is submitted to the protocol layer after being submitted. The protocol layer contains smart contracts and consensus mechanisms under block chains, providing contract creation and execution, consensus algorithm, and information verification. The protocol layer sets intelligent conditions such as network subscriber rules contract, information specification, user behavior specification, screening and tagning network user behavior and information. After screening, the data encryption layer is entered, and the network information is encrypted with the corresponding user in accordance with the data encryption algorithm, and the security and non-tamper-making of the guaranteed data storage. After the data information is processed, it is stored on each block in the block chain environment in the form of distributed storage. Distributed storage improves data storage space to achieve detrimentary data storage and application.

Under the blockchain technology, the network user behavior management system, on the premise of ensuring the normal network behavior of network users, realizes the constraint and analysis of network user behavior by creating intelligent contract and consensus mechanism algorithm, filters non-standard information, and marks dishonest network users. On the basis of purifying the network environment, the user information and data are encrypted and distributed stored to improve the security of data and the reliability of the system platform.

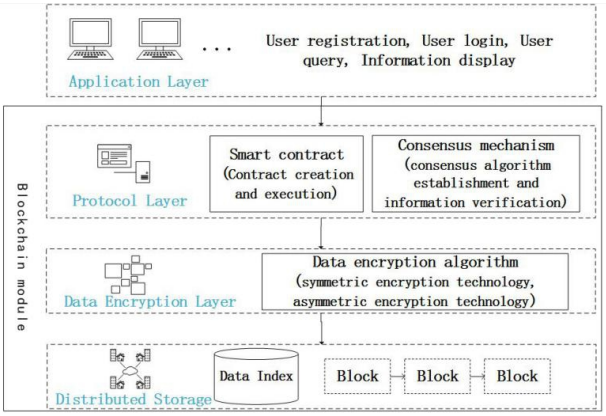


FIGURE 1. Figure 1 Framework of network user behavior management system based on blockchain Technology

IV. DETAILED DESIGN OF NETWORK USER BEHAVIOR MANAGEMENT SYSTEM BASED ON BLOCKCHAIN TECHNOLIGY

A . Process design of network user behavior management system based on blockchain technology

In the network user behavior management system based on blockchain technology, the user behavior of network users after information registration is constrained by relevant protocols. The system protocol layer can create corresponding contracts and mechanisms to control network user behavior. The specific process is shown in Figure 2.

After the user enters the system, submit the behavior application, and the user behavior application is activated and enters the protocol layer. The protocol layer is formed by intelligent contract and consensus mechanism algorithm. System administrators first create basic agreements such as laws and regulations, platform constraints, and create custom intelligent contracts in real time. After the consensus has reached a consensus, the contract takes effect. After the user's behavior information is incorporated into the protocol layer, the legality of the information is first verified. Verify invalid information will perform the rejection user behavior, information tag

and black, and mark user informal behavior. This ensures the legality of user behavior in the network environment. If the verification information is legal, the user behavior is executed and recorded. According to the security level of legal user behavior information, the system is encrypted. The system encrypts data information based on the hash function, digital signature technology, timestamp tag. In response to the implementation and storage of legitimate information, the intelligent algorithm can be analyzed. System analysis network user behavior track, predictive behavior development, based on analysis data, establish more advanced intelligent contracts, improve system security and intelligence. The encrypted data information stores in the form of a decentralized block chain environment in the form of distributed storage. The system combines storage data such as user name, timestamp, information keyword.

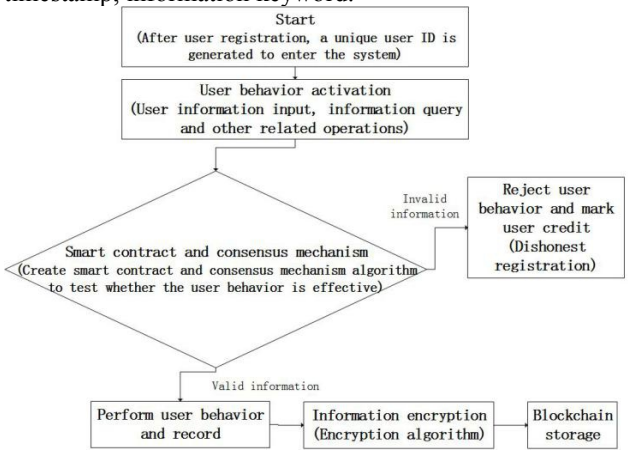


FIGURE 2. Figure 2 Flow chart of network user behavior management system based on blockchain Technology

B . Data flow design of network user behavior management system based on blockchain technology

Under block chain technology, network user behavior management systems form a standardized data stream from user registration. Data streams will change the data item content in the system layers, and increase the corresponding operation information until the system flow is completed. The entire system data stream is shown below (Figure 3):

- The user uses the system through the browser or client in the application layer. After registration in the system, the unique username UI, the UI is an abbreviation for the unique user ID, including the personal information required for user registration;
- Users use the UI account to enter the system, submit a behavior operation application (including the upload, information update, information query, information deletion, information extraction, etc.). According to the user behavior instruction, the application time (timestamp) and UI, the system finally generates user behavior registration data items: (BI, TS, UI). Among them, BI is an abbreviation of behavioral instructions, TS is an abbreviation of timestamp;
- (BI, TS, UI) data item enters the protocol layer, performs verification of intelligent contract and consensus mechanism. Intelligent contracts include the created laws and regulations contract FC, user rights protection contract BH, user credit management contract XC, personalized contract BS, etc. The consensus mechanism includes a working volume certificate mechanism (POW), equity certification mechanism (POS), part of the byzantine consensus algorithm (PBFT) and other consensus algorithms. According to the intelligent contract and consensus mechanism, the generated BSn(n is the intelligent contract number), and the PSn (n is a consensus algorithm sequence number), and finally in the form of (BI, TS, UI, BSn, PSn) data items. If the data in the (BI, TS, UI) data items do not meet the protocol layer, the username UI is marked, and the operation request for the (BI, TS, UI) data item is denoted, thereby logging the corresponding illegal information and stored ;
- (BI, TS, UI, BSN, PSN) data item according to data encryption, according to the encryption algorithm generating ciphertext ASR (cipheet ASR mainly according to the user name UI private key, the encryption algorithm rule generates the corresponding public key, and then with the help existing public key encryption forms an ASR ciphertext);
- Data ciphertext ASR incoming block chain distributed data store, according to the detrimentary characteristics of the block chain, the data dispersion stores and the corresponding blocks to facilitate the operation of the data.

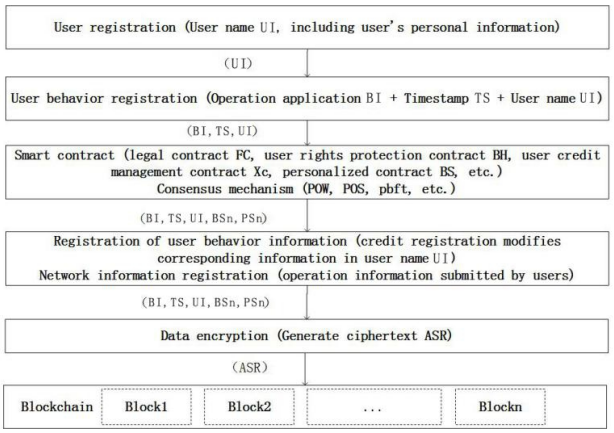


FIGURE 3. Figure 3 Data flow diagram of network user behavior management system based on blockchain Technology

BSN and PSN will mark the agreement execution degree corresponding to the user's behavior and record the user's credit degree. For different contract users' breach of contract, non-compliance with contract provisions and violation of protocol rules, the user's credit degree will be reduced to a certain value. The system will prohibit users from publishing information or extracting data, thus restricting the network user's behavior. The smart contract contains the required contract terms. The basic terms can be set, including legitimacy constraints, user rights and interests protection, or user-defined terms, including trusted operation, increasing credit, etc., to intelligently

standardize the system platform environment. The consensus mechanism can set up a common permission mechanism according to the consensus algorithm and workload to reach the consensus of all users.

Every operation of the data in the system will be recorded and stored in the corresponding database. The data can not be tampered with in the whole system process. The process and log will be distributed storage and block chain environment according to the needs to ensure the irreversibility and security of the data, and provide strong support for the authenticity of the system data.

C. *Internet user behavior management system experiment and analysis*

Based on zone chain technology, this paper builds a network user behavior management system, providing a new technology platform for purifying network information environments, protecting user behavior privacy, and improving the communication rate of network platform information. The experiment of network user behavior management system under block chain technology is mainly designed with the B / S structure (Browser / Server, browser / server mode), and the experiment selection Windows 10 operating system, i5-9500 CPU, 8GB memory. System development selection SOLIDITY programming language, Visual Studio Code software is a development environment and uses assistive development of HTML, JS.

This paper climbs 2000 network resource data from the academic forum, such as small wooden worms, generic passengers, after screening, preliminary treatment, reserved 1200 effective information data as the experimental data group A (N) of this article, A (N) enters the user After the behavioral management system, the information tag is first tagged according to the user released, including information such as the user's nickname, account registration, and the labeled experiment data group AT (n). A-T (N) has passed the intelligent contract layer, the user behavioral information registration layer, the data encryption layer, and finally demonstrates and saves the experimental data group according to the user behavior information flow. Since the experimental data group A (N) is before entering the system, this paper has been artificial screening, including the validity, accuracy and availability of information.

Table 1 shows some of the results of this paper system experiment. Experiments have proven that entering 1200 effective information, the system will perform username marks, contract screening and registration, user behavior information protection, etc., the resulting results are compared with the results of the artificial screening, and can see the system's implementation accuracy. , effective information rate, etc., the information rate after manual screening is small, and the encrypted storage and propagation of information has a high execution effect.

The experimental results can be obtained, the network user behavior management system has high information screening, user identification and information filtering capability, can mark network user information, information validity screening, intelligent contract provisions user comprehensive behavior information matches, and finally encrypts user information. In the above table, the comparison between the system and the artificial screening is shown. It can be seen that this article has a high execution effect, but the result of the organizational screening, also has a certain rise in rising space, and it also requires re-accuracy. Implementation results.

TABLE I. THIS ARTICLE SYSTEM EXPERIMENT DATA EXECUTION RESULT IS COMPARED

Experiment array A(n)	***User information tag***	***Information valid screening***	***Terms of contract***	***User behavior match***	***Encryption***
Perform success rate / this article	98.4%	95.1%	98.1%	98.8%	98.0%
Data effective / artificial screening	97.6%	99.3%	98.9%	99.2%	99.4%

V . SUMMARY AND PROSPECT

Under the blockchain technology, the network user behavior management system takes the application layer, protocol layer, data encryption layer and distributed storage layer as the main architecture mode. With the help of blockchain intelligent contract, consensus mechanism, data encryption, distributed storage and other sub technologies, the complex network user behavior under the Internet is intelligently managed and distributed screened in the form of contract and protocol encryption algorithm as the core to ensure the security of data, and on the basis of data flow registration and storage to ensure that the data can not be tampered with, for the Internet era to create intelligent, standardized network user behavior management system, purify the Internet environment.

This paper puts forward the theoretical framework and data flow design process of network user behavior management system based on blockchain technology, but the detailed design of smart contract, consensus mechanism and data encryption has not been put forward specific analysis. In the future, we will carry out detailed research and problem optimization on Intelligent contract, consensus mechanism and data encryption sub technology, so as to better improve the intelligence and security of the system.

REFERENCES

[1] Pan Juncheng, Deng Suping. Reconstruction of science and technology relationship based on blockchain technology in peer review of journals [J]. Publishing wide angle, 2020 (06): 50-52 + 70

[2] Lin Haohan. Research on network rumor governance based on blockchain technology [D]. Inner Mongolia University, 2019

[3] Zhou Tong. Research and application of trusted data token based on blockchain technology [D]. University of science and technology of China, 2019

[4] Wang Xiuli, Jiang Xiaozhou, Li Yang. Data access control and sharing model based on blockchain [J]. Acta software Sinica, 2019,30 (06): 1661-1669

[5] Ma Lichuan. Research on trust management mechanism in group intelligence collaborative network [D]. Xi'an University of Electronic Science and technology, 2018

Research on Smart Home Security Threat Modeling based on STRIDE-IAHP-BN

1st Rongjuan Zhu
School of Artifical Intelligence and Computer Science
Jiangnan University
Wuxi,China
zrj15161@163.com

2nd Xinke Wu
School of Artifical Intelligence and Computer Science
Jiangnan University
Wuxi,China
6191914012@stu.jiangnan.edu.cn

3rd Jun Sun
School of Artifical Intelligence and Computer Science
Jiangnan University
Wuxi,China
junsun@jiangnan.edu.cn

4th Zhihua Li
School of Artifical Intelligence and Computer Science
Jiangnan University
Wuxi,China
w*xzhli@aliyun.com*

***Abstract* - As a common form of Internet of things applications, smart home also faces security threats. Because majority of the network security threat modeling are Internet-oriented. It is a challenging and scientific question to how to treat modeling of smart home. This paper combines the improved analytic hierarchy process with Bayesian network and proposes a method of STRIDE-IAHP-BN for threat quantification and priority definition. Moreover, in order to verify the effectiveness of the method, the DVR system is selected to analyze and verify the method. The results of simulation show that the modeling method can provide effective support for the assessment of smart home security threat.**

Key words: security threat modeling; the method of STRIDE-IAHP-BN; improved analytic hierarchy process; smart home

I. INTRODUCTION

The Internet of Things (IoT) is one of the national key strategic emerging industries recognized by the world with broad development prospects [1]. With the rapid development of IoT, how to improve the security of IoT has become a common challenge for the academic and industrial fields [2]. The hidden danger of IoT not only brings property loss to users, threatens their life security, and even poses a great threat to the security of the country's major infrastructure. Countries all over the world have invested a great deal of money to conduct in-depth research in the field of IoT information security and the security of IoT applications [3]. As a common form of IoT applications, smart home has advantages in low cost, powerful functions and convenient deployment, but there are major security risks such as being vulnerable to remote network attacks and privacy disclosure [4]. Network security threat modeling is a basic technology and method for studying and evaluating network security risks, which can provide a powerful reference for situation awareness and network risk early warning and prevention [5]. Therefore, how to model the security threats of smart home is a challenging and scientifically significant research topic.

At present, research on IoT threat modeling is still in the initial stage. In 1999, Bruce Scheier [6] defined attack tree for the first time. Attack tree is mainly used for description and formal analysis, which has the disadvantage of strong subjectivity. Based on the concept of attack tree, scholars [7] proposed a threat modeling method for IoT, which adopts a strategy similar to brainstorming in the model implementation. Unfortunately, this method is not suitable for specific assessment work, nor is it suitable for large-scale network systems with high

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00059

complexity; Microsoft has proposed a STRIDE model, which mainly draws data flow diagrams and obtains threat indicators based on the Microsoft Threat Modeling Tool. However, this model cannot fully express the security concept, data elements, abstraction level, deployment information and simplified system [8]. Zhou Yan [9] proposed an improved STRIDE threat modeling method, in which the dimensional information of the location where the threat occurred was added on the basis of stride, but the critical factors such as severity and priority of threat were not fully considered. In addition, there are some common threat modeling models, such as Common Vulnerability Scoring System (CVSS) [10] and DREAD rating system [11], but they are not intuitive enough for threat modeling, and the measurement basis that the rating relies on is also very general, and it has high professional requirements for the raters. Sun Ao et al. [12] proposed a STRIDE-HMM risk assessment model and method combining STRIDE and Hidden Markov Model (HMM). This method carries out quantitative calculation of threats, but the state transition probability matrix and observation probability matrix are given by subjective experience.

In this paper, STRIDE-IAHP-BN, a comprehensive quantitative modeling method for threat quantification and priority definition is proposed based on the Improved Analytic Hierarchy Process (IAHP) combined with Bayesian Network (BN). The STRIDE-IAHP-BN method scores based on the index risk value (high, median or low) obtained by STRIDE and the number of the third-level indexes belonging to the second-level indexes, which can reduce the influence of subjective factors to a certain extent; furthermore, the consistency problem is transformed into a constrained programming problem and solved by CPLEX Optimizer; finally, the probability obtained by IAHP is taken as the prior probability of BN, and combined with the joint probability distribution table of BN, the final risk probability of the threat is calculated. In this way, security personnel can take corresponding countermeasures more accurately based on the probabilistic priority and severity of threats. In addition, in order to verify the effectiveness of the model, the DVR system is selected to analyze and verify the model. The experimental results show that the model can provide effective decision support for threat assessment of smart home security.

The rest chapters are arranged as follows: The first chapter mainly introduces IAHP and the proposed modeling method of Stride-IAHP-BN; the second chapter mainly introduces the threat modeling method based on Stride-IAHP-BN proposed in this paper and the example demonstration in the DVR system, and finally, the conclusion is drawn.

II. STRIDE-IAHP-BN MODEL

A. Improved Analytic Hierarchy Process

In order to make the initial weights of the second and third level index nodes more objective and maneuverable, the analytic hierarchy process is improved in this paper. Firstly, the acquisition of traditional judgment Matrix relies on the experience and knowledge of experts, which is relatively subjective. In this paper, we adopt the strategy of scoring the quantity of third-level indexes to which second-level indexes belong according to STRIDE's analysis, the more the number, the higher the score; the third-level indexes are scored according to the risk level of high, median, and low. The higher the risk level, the higher the score. the benefits of doing so is that the difficulty and subjectivity of scoring are reduced to a certain extent; secondly, in calculating the weights, the traditional AHP methods such as geometric mean, arithmetic mean, eigenvector, least squares and so on are avoided, instead, the consistency problem is transformed into a mathematical programming problem, which effectively circumvents the requirement of consistency; finally, the mainstream solver CPLEX and YALMIP are imported into MATLAB R2016b, and the judgment Matrix is weighted by least squares. The objective function and constraints are shown in formula 6. Based on the above improvements, an improved AHP (IAHP) method is proposed.

$$\min Z = \sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij}w_j - w_i)^2 \qquad (6)$$

$$s.t.\sum_{i=1}^{n} w_i = 1(w_i > 0, i = 1,2,...,n)$$

where i=1,2…n; ω_i is the weight of evaluation factor i; n is the number of evaluation factors and a_{ij} is the judgment matrix.

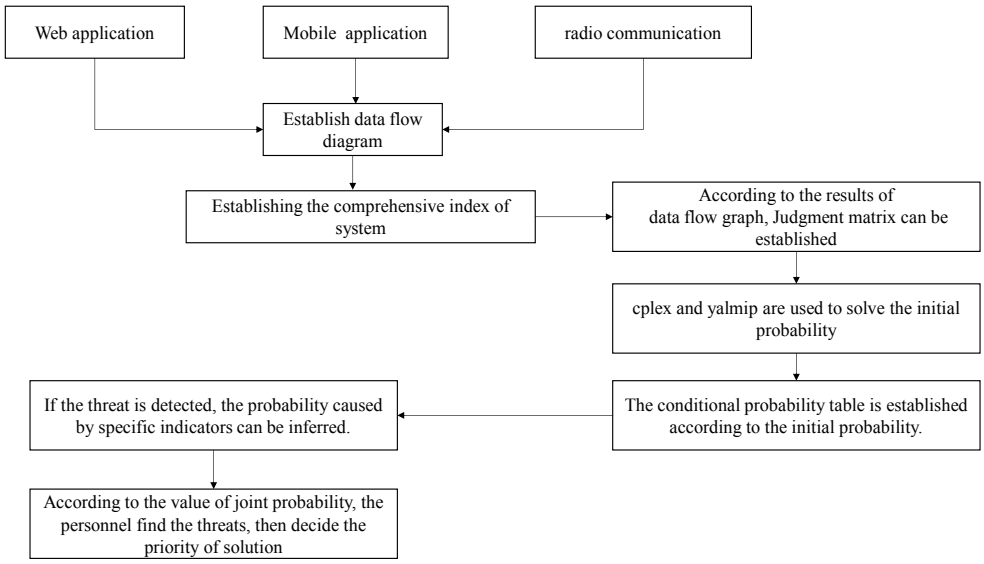
The IAHP algorithm is described as follows:

```
Input: N*N      //* judgement matrix
Output:1*N      //* weight of index
       C = [
       x>0
       sum(x)
       ]                    //*constraint condition
       for i=1:N
            for j=1:N
                 z=z+(n(i,j))*x(j)-x(i))^2   //*objective
function
            end
       end
       if result.problem == 0    //* success
       else
            failure
end
```

The time cost of IAHP mainly comes from the double-loop statement to solve the objective function, so the complexity of the algorithm is $0(N^2)$. The improved analytic hierarchy process makes the calculation of index weight easier and quicker by improving the procedure of judging consistency, thus speeds up the convergence speed of the algorithm.

B. Construction Method of IAHP Threat Assessment Model

With the quantitative assessment of smart home security threats as the demand, the following discussion is made on the construction process of threat assessment model. Firstly, based on the Threat Modeling Tool, the data flow diagram is constructed from the three aspects of web application, mobile application and radio communication, and the smart home security threat index system is constructed, and all the threat indexes are divided into three levels: high, median, and low； secondly, the proposed IAHP method is used to calculate the weight of indexes, and the weights obtained are used as the initial probabilities of the second-level and third-level index nodes; finally, the joint probability that the final security risk may occur can be calculated according to the Bayesian inference model by setting joint probability table of nodes. Thus, based on this joint probability, security practitioners can conduct threat detection, threat early warning, and determine the priorities of response measures. It can be seen that the method is universal to a certain extent. To sum up, a model for quantitative security threat assessment (STRIDE-IAHP-BN) is proposed, as shown in Figure 1.



III. APPLICATION OF STRIDE-IAHP-BN MODEL

The number of stencils available in the Threat Modeling Tool for representing devices, communication transfers, and input/output trusted boundaries is limited, so you need to create your own stencils during the actual drawing process. Figure 2 shows the interface for adding a stencil by yourself; Figure 3 shows the Web application

data flow diagram of DVR system; Figure 4 shows the mobile application data flow diagram of DVR system; Figure 5 shows the radio communication data flow diagram of DVR system.

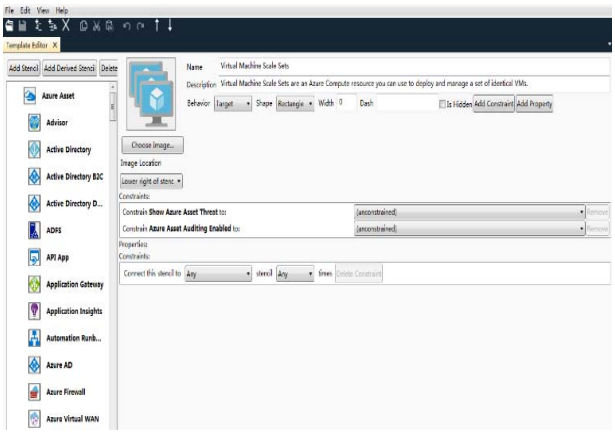


Figure 2. Interface for adding a stencil by yourself

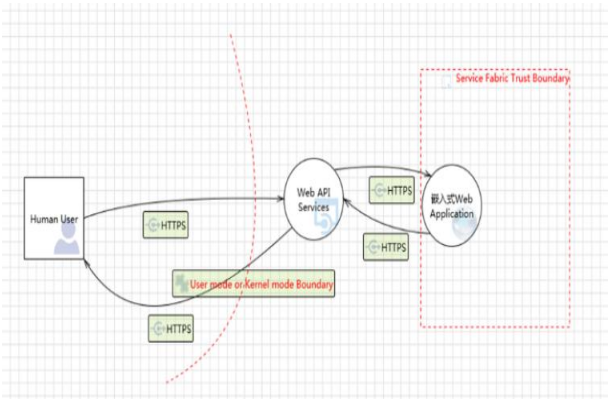


Figure 3. Web application data flow diagram of DVR system

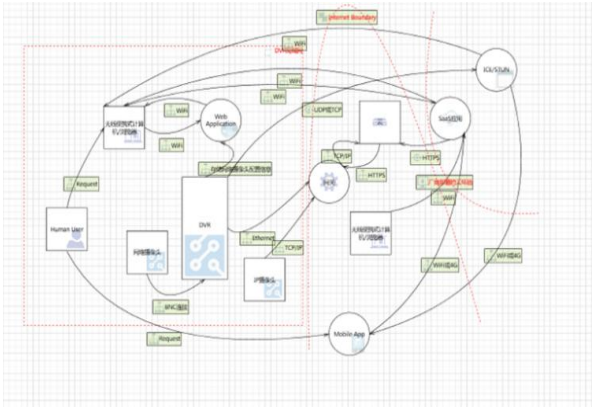


Figure 4. mobile application data flow diagram of DVR system

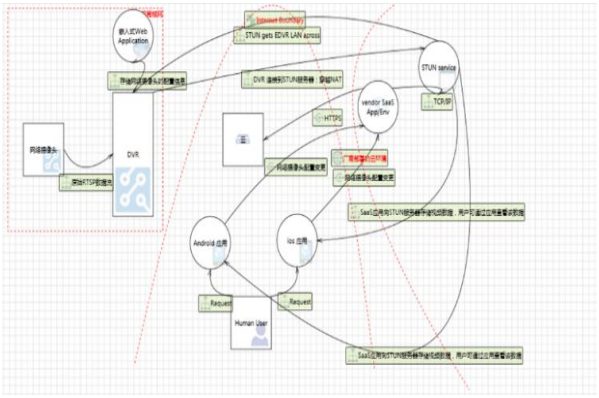


Figure 5. radio communication data flow diagram of DVR system

Furthermore, taking the Web application data flow diagram of DVR system as an example, the threat analysis diagram shown in Figure 6 can be generated by running the data flow diagram.

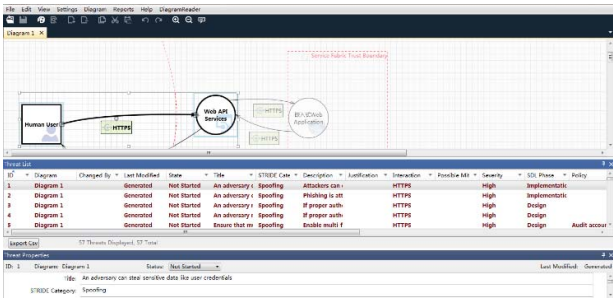


Figure 6. Threat analysis diagram

Finally, a threat analysis report as shown in Figure 7 is generated by Microsoft Threat Modeling Tool.

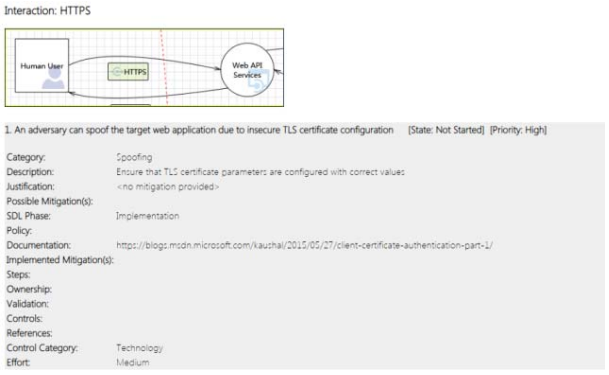


Figure 7. Threat analysis report

Through the analysis of the report shown in Figure 7, it is found that there is a total of 50 threats in all links. The quantity of each threat type is shown in Table 1.

Table 2. Quantity of each threat type

Threat type	Quantity
S	14
T	12

R	3
I	10
D	2
E	9

The identified threat examples are described into a document, which contains the description of the threat, the target of the threat, the attack technology, countermeasures, and the risk level.

The traditional AHP is subjective in scoring and the initial judgment Matrix usually does not meet the requirements of consistency, which need to be adjusted several times, therefore, in this paper, the IAHP method proposed is adopted to calculate the weight of indexes, and the STRIDE method is used, in addition, the consistency problem is transformed into a constrained programming problem, so as to effectively overcome the above shortcomings. As for the scoring method of the B-C judgment matrix between the index level and the criterion level, if the two indexes have the same risk, i.e. both of them are high risk, median risk or low risk, the score is 1; if the risk values of the two indexes are respectively one high, one median or one median, one low, the score is 5 or 3, and one high, one low, the score is 9 or 7. This method greatly reduces the influence of subjectivity.

Finally, after the evaluation is completed, the corresponding judgment matrix is input into the CPLEX Optimizer to obtain the weight of each index. Here, only the weights of prior probabilities of indexes required by the following Bayesian network are listed: 0.0673 for B3, 0.0192 for C27, 0.0388 for C28, and 0.0094 for C29.

The relationship between the number of nodes n in the Bayesian network structure diagram and the joint probability distribution table is 2^n. To carry out the Bayesian network precise reasoning shown in Figure 7, we need 2^{57} combined conditional probability tables [23]. Therefore, only B3 (denial) branch is selected for discussion. Convert C27 (an attacker can deny access to the cloud gateway due to lack of audit) , C28 (an attacker can deny the API malicious behavior on the API that caused the denial problem) , and C29 (an attacker can reject the attack footprint of a malicious problem) to Bayesian root nodes, and convert event B3 into a leaf node of Bayesian Network. Figure 9 shows the Bayesian network model of this branch.

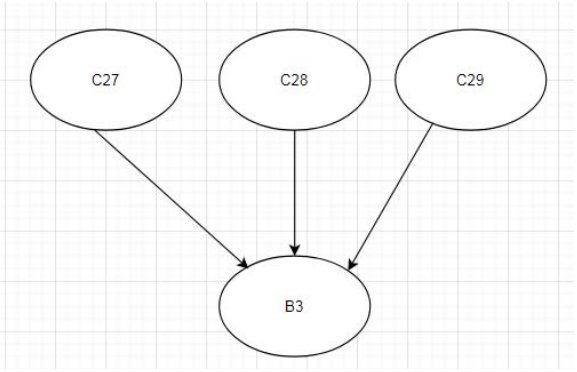


Figure 9. Bayesian network model

Further, the weights obtained in the parameter learning stage based on AHP are set as the prior probability of each node, and the conditional probability table is set. As shown in Figure 10, 1 corresponds to Fault, that is, the probability of a threat, and 0 corresponds to OK, that is, probability that the system is safe without threat.

C27	P(C27)
1	0.0192
0	0.9808

C28	P(C28)
1	0.0388
0	0.9612

C29	P(C29)
1	0.0094
0	0.9906

C27	C28	C29	P(B3)
0	0	0	OK
0	0	1	Fault
0	1	0	Fault
0	1	1	Fault
1	0	0	Fault
1	0	1	Fault
1	1	0	Fault
1	1	1	Fault

Figure 10. Prior/conditional probability table of Bayesian network

Table 9. Joint probability of Bayesian network

C27	C28	C29	B3	P(C27)	P(C28)	P(C29)	P(B3|C27,C28,C29)	Product
OK	OK	OK	OK	0.9808	0.9612	0.9906	1	0.933883
OK	OK	Fault	OK	0.9808	0.9612	0.0094	0	0
OK	Fault	OK	OK	0.9808	0.0388	0.9906	0	0
OK	Fault	Fault	OK	0.9808	0.0388	0.0094	0	0
Fault	OK	OK	OK	0.0192	0.9612	0.9906	0	0
Fault	OK	Fault	OK	0.0192	0.9612	0.0094	0	0
Fault	Fault	OK	OK	0.0192	0.0388	0.9906	0	0
Fault	Fault	Fault	OK	0.0192	0.0388	0.0094	0	0
OK	OK	OK	Fault	0.9808	0.9612	0.9906	0	0
OK	OK	Fault	Fault	0.9808	0.9612	0.0094	1	0.008862
OK	Fault	OK	Fault	0.9808	0.0388	0.9906	1	0.037697
OK	Fault	Fault	Fault	0.9808	0.0388	0.0094	1	0.000358
Fault	OK	OK	Fault	0.0192	0.9612	0.9906	1	0.018282
Fault	OK	Fault	Fault	0.0192	0.9612	0.0094	1	0.000173
Fault	Fault	OK	Fault	0.0192	0.0388	0.9906	1	0.000737
Fault	Fault	Fault	Fault	0.0192	0.0388	0.0094	1	0.000007

It is not difficult to find from Table 9 that if a threat is detected, that is, if the detection result is Fault, then the result is the probability caused by C27, which is calculated as follows:

$$P(C27 = \text{Fault}|D = \text{Fault}) = \frac{0.18282+0.000173+0.000737+0.000007}{0.008862+0.037697+0.000358+0.018282+0.000173+0.000737+0.000007} = 29.0\%$$

. Similarly, if a threat is detected in the system, but no C28 threat is found after investigation, then under this condition, the probability of C27 occurrence is as follows:

$$P(C27 = \text{Fault}|B3 = \text{Fault}, C28 = OK) = \frac{0.0128282+0.000737}{0.008862+0.018282+0.000173} = 70.0\%$$

. Obviously, the security personnel may refer to the probability value for threat detection and response.

IV. CONCLUSION

In this paper, based on the method of STRIDE, a quantitative method combining the improved analytic hierarchy process and Bayesian network, which can reduce the demand for prior knowledge of the scorer and the difficulty of analysis compared with the CVSS scoring system and the DREAD rating system, is proposed; the improved analytic hierarchy process reduces the scale, and the risk level and the quantity of indexes obtained by STRIDE greatly reduce the subjectivity of scoring; in addition, the consistency problem is circumvented by turning it into a constraint programming problem. The prior probabilities in the Bayesian network are derived from an improved analytic hierarchy process, which is more accurate than the prior probabilities based on experience. The Bayesian Model replaces the Hidden Markov Model and provides a more accurate fine-grained model. Once the risk probability of the index reaches a certain threshold, it is necessary for the security personnel to take security policies for adjustment. In addition, measures to mitigate the threat are provided. This threat modeling method can provide security personnel with a basis for decision-making, so as to build a safe and reliable defense.

REFERENCES

[1]. Wu H, Han H, Wang X, et al. Research on Artificial Intelligence Enhancing Internet of Things Security: A Survey[J]. IEEE Access, 2020, 8: 153826-153848.

[2]. Mishra P, Biswal A, Garg S, et al. Software Defined Internet of Things Security: Properties, State of the Art, and Future

Research[J]. IEEE Wireless Communications, 2020, 27(3): 10-16.
[3]. [American] Brian•Russell Drew•van•durren By. Internet of things security[M]. translated by Li Wei Shen Xin Hou Jingyi Wang Ziliang Beijing: China Machine Press,2018.
[4]. Agrawal D, Bhagwat R, Bandopadhyay R, et al. Enhancing Smart Home Security using Co-Monitoring of IoT Devices[C]//Companion of the 2020 ACM International Conference on Supporting Group Work. 2020: 99-102.
[5]. Li, Zhang,Xin, Tong. Threat Modeling and Countermeasures Study for the Internet of Things[J]. Journal of Convergence Information Technology,2013,8(5).
[6]. Dewri R, Poolsappasit N, Ray I, et al. Optimal security hardening using multi-objective optimization on attack tree models of networks[C]//Proceedings of the 14th ACM conference on Computer and communications security. 2007: 204-213.
[7]. Lallie H S, Debattista K, Bal J. A review of attack graph and attack tree visual syntax in cyber security[J]. Computer Science Review, 2020, 35: 100219.
[8]. Stone E E, Skubic M. Passive in-home measurement of stride-to-stride gait variability comparing vision and Kinect sensing[C]//2011 Annual international conference of the IEEE engineering in medicine and biology society. IEEE, 2011: 6491-6494.
[9]. Zhou Yan. Research on improved STRIDE threating model[D].Wuhan: Huazhong University of Science and Technology,2015.
[10]. Mell P, Scarfone K, Romanosky S. Common vulnerability scoring system[J]. IEEE Security & Privacy, 2006, 4(6): 85-89.
[11]. Burke M J, Salvador R O, Smith-Crowe K, et al. The dread factor: how hazards and safety training influence learning and performance[J]. Journal of Applied Psychology, 2011, 96(1): 46.
[12]. Sun Ao, Yin Xiaochuan, Li Xiaoqing. A task oriented network risk assessment model[J]. Journal of Air Force Engineering University,2020,20(5):105-110.
[13]. Danielis P, Beckmann M, Skodzik J. An ISO-Compliant Test Procedure for Technical Risk Analyses of IoT Systems Based on STRIDE[C]//2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC). IEEE, 2020: 499-504.
[14]. Ghosal A, Halder S, Conti M. STRIDE: Scalable and Secure Over-The-Air Software Update Scheme for Autonomous Vehicles[C]//ICC 2020-2020 IEEE International Conference on Communications (ICC). IEEE, 2020: 1-6.
[15]. Saaty T L. Decision making with the analytic hierarchy process[J]. International journal of services sciences, 2008, 1(1): 83-98.
[16]. Saaty T L. What is the analytic hierarchy process[M]//Mathematical models for decision support. Springer, Berlin, Heidelberg, 1988: 109-121.
[17]. Hosseini S, Ivanov D, Dolgui A. Ripple effect modelling of supplier disruption: integrated Markov chain and dynamic Bayesian network approach[J]. International Journal of Production Research, 2020, 58(11): 3284-3303.
[18]. Boutkhamouine B, Roux H, Pérés F. Data‐driven model for river flood forecasting based on a Bayesian network approach[J]. Journal of Contingencies and Crisis Management, 2020, 28(3): 215-227.
[19]. [American]By Adam Stark.threating model[M]. translated by Jiang Changqing,and so on.Beijing: China Machine Press:2015.
[20]. By [American] Alan • guzman Aditia • Gupta. Internet of things penetration test[M].tanslated by Wang Bin.Beijing: China Machine Press,2019.
[21]. Sion L, Yskout K, Van Landuyt D, et al. Security Threat Modeling: Are Data Flow Diagrams Enough? [C]//IEEE/ACM 42nd International Conference on Software EngineeringWorkshops (ICSEW'20). IEEE/ACM, 2020.
[22]. Li M, Wang H, Wang D, et al. Risk assessment of gas explosion in coal mines based on fuzzy AHP and bayesian network[J]. Process Safety and Environmental Protection, 2020, 135: 207-218.
[23]. He Yongchang,Chen Zhiguang,Wang haifeng and so on. Research on Bayesian network model of Missile Fault Diagnosis Based on netica[J]. Aviation Weapon,2020,27(1):89-95.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

A High Precision Refresh Method to Improve The Performance of Flash Storage Devices

Peixuan Li[1,2,3,4], Yaofang Zhang[1,2,3,4], Deli Yin[5], Ping Xie[1,2,3,4*]

1.College of Computer of Qinghai Normal University, Xining 810016, P. R. China
2.The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Xining 810016, P. R. China
3.The Key Laboratory of Internet of Things of Qinghai Province, Xining 810016, P. R. China
4.Academy of Plateau Science and Sustainability, Xining 810016, P. R. China
5.Chuzhou Polytechnic, Chuzhou, 239000, P. R. China
**Corresponding author: xieping@qhnu.edu.cn*

***Abstract*—With the development of flash memory, its storage density has gradually increased, but its reliability has been greatly reduced. Now, flash memory relies on LDPC to solve this problem. Although LDPC (Low Density Parity Check Code) can correct data with a high error rate, it will cause a lot of delays. The refresh operation can cause the LDPC decoding delay. In the work of this article, we propose a precise refresh strategy. On the basis of high latency pages, according to the characteristics of data access frequency, we take the read-hot data in the write-cold data as our refresh target. We test the performance of the precise refresh scheme through different workloads. Compared with ordinary refresh schemes, the average response time of flash memory storage devices is reduced by 16%-49% and the lifetime of flash memory is extended.**

Keywords-LDPC, refresh, Nand flash, hot data

I. INTRODUCTION

Nowadays, SSD (Solid State Drive) with flash memory as storage media are very popular in computer systems and mobile devices. It has the advantages of fast reading speed, small size, lightweight, etc. In consumer products, MLC/TLC (multi-level/three-level cell) flash memory chips that store 2 or 3 bits per cell are used[1][2]. Although the cost price is gradually decreasing but the capacity is gradually increasing. And the storage density is higher. As storage density increases, devices need more precise voltage control capabilities when reading data. In addition, because multiple bits are compressed into one unit, the fault tolerance of the flash memory is lower, and voltage changes are more likely to cause errors in the flash memory. The improvement of the error rate of flash memory makes the error correction capability of the traditional BCH (Bose Chaudhuri Hocquenghem) error correction code no longer suitable for solid state drives composed of high-density flash memory. Nowadays, the design of LDPC in solid-state hard drives is continuously being applied in practice[3][4].

LDPC provides a strong reliability guarantee for flash storage devices. LDPC uses an iterative way to decode. The purpose of decoding iteration is to accumulate bit decision information. The more iterations, the higher the reliability of the bit decision. It can be used to correct higher error rate data, but it will also cause a lot of error correction delays. Due to the current situation of high latency, two optimization methods are now proposed. The first way is to speed up LDPC decoding. It reduces the decoding waiting time by caching all kinds of LDPC information[5] or providing additional decision information[6] to the decoding process. The second is to reduce the data error rate. The method is to correct the error of the data and write to a new location when the error rate of the data exceeds a certain security level[7]. Although the data error rate is reduced, it has serious shortcomings. Because it will cause additional reading, programming, and erasing operations, thereby affecting the performance and lifetime of flash memory storage.

In order to reduce the negative impact of LDPC and refresh, this work proposes a precise refresh plan. The basic idea is that the refreshed framework uses periodic refresh to refresh high-latency pages that have write-cold attributes and read-hot attributes. Therefore, we need to analyze the attributes of the flash memory pages and define high-latency data. According to previous research, the reading level of LDPC is 1-7[8]. The higher the reading level, the higher the decoding accuracy and the greater the delay. We found that when the reading level of LDPC is higher than a certain threshold, its decoding delay increases significantly. In a refresh cycle, if a logical data page is written less than twice, it is a write-cold data page if it is read more than twice, it is a read-hot data page. The above scheme is implemented in a well-configured flash simulator. Experiments with several actual workloads show that the refreshed framework can improve the average response time of the device while ensuring the lifetime of the flash memory.

II. BACKGROUND AND MOTIVATION

A. LDPC Encode and Decode

When executing a read request, the SSD controller always performs LDPC decoding on the flash memory first, and then returns the error-corrected data to the host. LDPC decoding always starts with hard decision decoding. When hard decision decoding fails, LDPC will perform soft decision decoding. Soft decision decoding has strong error correction capability and fast convergence speed, but it has a large number of calculations and high hardware overheads. Because soft decision decoding will go through three steps[9]:

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00060

1) Reread operation: The reread operation reads one or more sets of data in the flash memory according to the way the soft data is constructed, which can obtain multiple sets of information useful for decoding.

2) Soft data construction: The reread operation will read out multiple groups of different information about flash memory data composed of "0" and "1". LDPC will map these information into soft data that can be decoded by LDPC decoding. Each bit of data has a LLR(Log Likelihood Ratio).

3) Perform soft decision decoding: After constructing the soft data, each LLR in the codeword is passed to the LDPC decoder. Each real number in the LLR represents the probability of each bit of data "0" and "1". A positive number represents a greater probability of "0", and a negative number represents a greater probability of "1". In this way, the mistake of flash memory can be corrected.

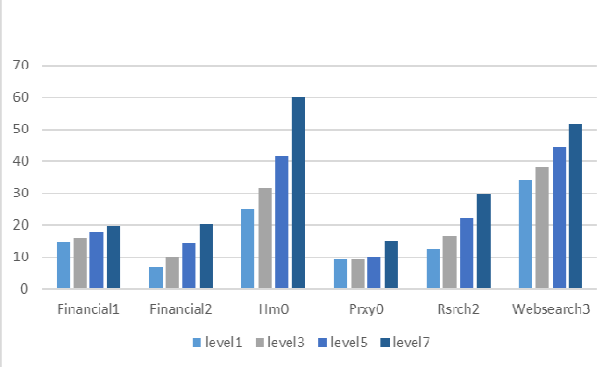


Figure 1. Refresh the framework

We use experiments to show the average access latency under different decoding levels, as shown in Figure 1. As the LDPC reading level increases, the average response delay of the device gradually increases. From level 5 to level 7, the average response delay of Financial2 increased by 39%, and Hm0 increased by 32%. When the LDPC reading level is 5, the average response time of the device has increased significantly, indicating that more LDPC decoding delays have occurred.

B. *Read-hot and Write-hot Data*

The hot or cold data is determined by the frequency of access. Read-hot data refers to data that has been accessed more frequently and read-cold data refers to data that has less access times. Identifying hot and cold data can effectively improve the performance of flash storage algorithms. Hot data recognition technology is used to reduce write amplification, improve garbage collection performance[10][11][12], etc. Not only read operations can affect data, but write operations also affect data. On different pages, the write frequency may be quite uneven. Some pages may be frequently written (write-hot pages), other pages are rarely updated (write-cold pages). The data written to the cold page is not updated frequently, and a large number of retention errors will occur, resulting in increased read latency. The write-hot page usually only has a low bit error rate. Even if the bit error rate is high, the page will be overwritten by new data due to the high update probability. The benefit of refreshing write-hot pages is low. So the benefit of refreshing cold pages is higher. Write-hot data only accounts for 1% of the total data[13], so accurate refresh still needs further analysis. At the same time, if a cold read page is refreshed, the system will no longer access the page. Then we only have the loss of refresh, but no gain. According to the popularity of the data, we propose a precise refresh plan.

III. ACCURATE REFRESH SCHEME

A. *Flash Refresh Framework*

In this section we introduced the design of the precise refresh framework, as shown in Figure 2. The address mapping module maintains the correspondence between the logical page address and the physical page address for the system. The refresh framework can get the basic information of all pages from the address mapping module. The ECC (Error Correcting Code) module provides decoding delay information for the refreshed frame. Determine whether it is a high-latency page according to the LDPC decoding level of the flash page. The refresh strategy proposed in this paper is to determine whether to refresh the high-latency page based on the write heat and read heat information of the flash memory page. The hot data judgment module will provide corresponding information for the refresh strategy. Every time interval, all data blocks that meet the refresh conditions will be refreshed to ensure data reliability. This scheme will be discussed in detail below.

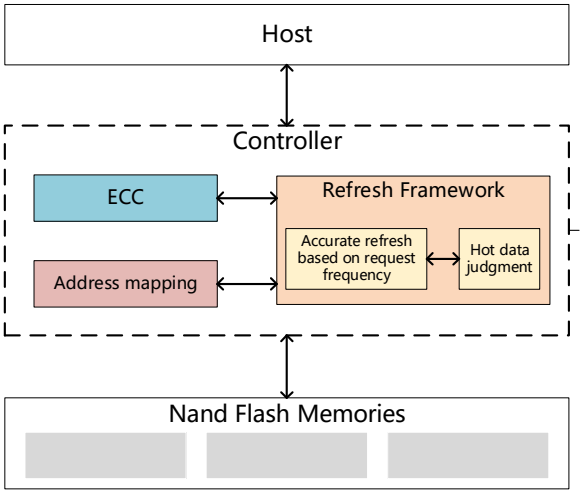


Figure 2. Refresh the framework

B. *Hot Data Judgment Module*

We have designed two queues, one is used to judge the hot read data page, the other is used to judge the hot write data page. When a data page is read for the second time, the address information of the flash memory page will be put into the hot read data queue. The hot data queue has limited storage space. If the hot read data has not been accessed for a long time, it will leave the hot read queue and become cold read data again. The same method can be used to determine the write-hot data page. After a cycle is refreshed, both queues will be emptied and the hot data will be judged again.

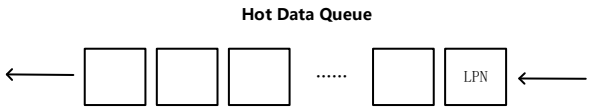


Figure 3. The process based on the high LDPC delay block refresh policy

C. *Accurate Refresh Based on Heat*

The accurate refresh based on heat scheme determines the high latency page to be refreshed according to the write heat and degree heat of the page, figure 4 shows the process of the high-precision refresh scheme. In a refresh cycle, the framework calculates the read frequency, write frequency, and LDPC error correction delay of flash memory pages. In the second chapter, we show through load experiments that the delay increases when LDPC decoding level 5 is greater. Therefore, pages with LDPC decoding levels above level 5 are high-latency pages. If the flash memory has been written (read) data more than twice, the page is read-hot (write-hot) data. If the data stored in a high-latency flash page is both write-cold data and read-hot data, the data page is refreshed. The heat information of the data is given by the heat data judgment module.

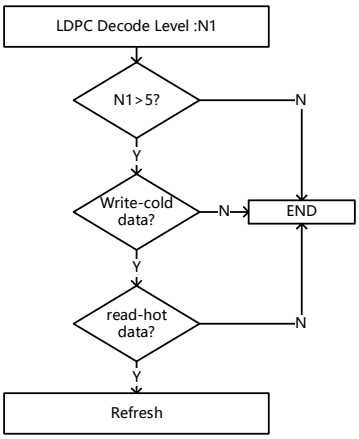


Figure 4. The number of grades above 3 level and above 5 level within the block mark the update process.

IV. EXPERIMENTTAL ANALYSIS

In this section, we show the basic configuration of the experimental device and flash memory storage device. Test the effect of refresh strategy through workload experiments. The impact of the refresh strategy on the average response time and lifespan of flash storage devices is analyzed.

A. *The Experimental Device*

We added Microsoft's SSD module to DiskSim[14] to simulate flash storage devices. The storage device is configured with 8 channels, each channel has 8 chips, each chip has 4 planes, each plane has 2048 blocks, each block has 64 8KB pages and the LDPC software decode read level is set to 1-7. In terms of workloads, eight workloads with different read/write ratios were used to test the effect of the refresh strategy based on high LDPC delay blocks. We listed the information for each load, as shown in Table 1.

TABLE I. WORKLOAD INFORMATION

Workload	Read	Write	Read ratio
Financial1	1235633	4099353	23%
Mds0	143973	1067061	12%
Prxy0	383524	12135444	3%
Hm0	1417748	2575568	37%
Rsrch2	136364	71223	65%
Financial2	3046112	653082	82%
Proj0	13006718	1907336	87%
Websearch3	16410267	26772	98%

During the experiment, we evaluated four strategies:

1) The original strategy: In the original strategy, the flash storage device does not have the refresh capability and only relies on the device's garbage collection module, ECC module and wear balance module to ensure the reliability of data..

2) Traditional refresh strategy: In the traditional refresh strategy, datas are refreshed only when the data error rate reaches the threshold that cannot be corrected by the ECC module.

3) The Accurate refresh strategy: This is the refresh strategy proposed in this article. Refresh high latency pages according to the characteristics of write heat and read heat.

B. *Experimental results and analysis*

In this section, we analyzed the average response time and the number of refreshes for precise refresh.

Figure 5 shows the normalized average response time of the three strategies. The precise refresh scheme reduces the average response time of the workload to a certain extent. The higher the read ratio, the more read-hot data. The optimization of precise refresh to read latency is more obvious. For loads with a reading ratio of more than 50%, their average response time is reduced by 16% to 49%.

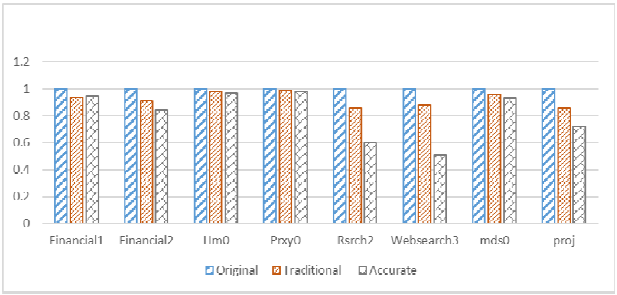


Figure 5. Normalized average response time

Average response delay optimization:

Figure 6 shows the normalized refresh times of the traditional refresh strategy and the precise refresh strategy. Compared with the traditional refresh strategy, the refresh volume of the precise refresh strategy is reduced by 12% to 58%. Although the average response time optimization effect of financial1, prxy0, and other workloads is not obvious. However, the number of refreshes is significantly reduced, and under the same conditions, the service life of the flash memory storage device is prolonged.

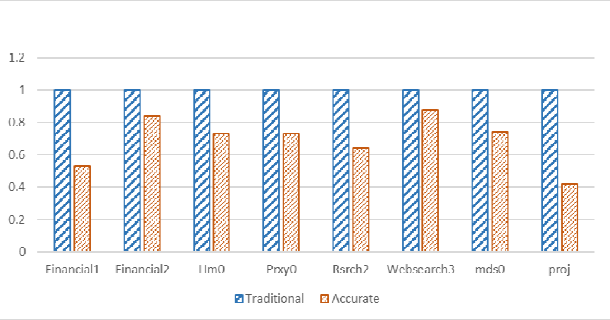


Figure 6. Normalized refresh times

V. CONCLUSION

With the advancement of storage technology and the increase in the density of flash memory devices, the reliability of flash memory has declined. Using data refresh operation can increase the reliability of flash memory and improve the read and write performance of flash memory storage devices. However, frequent refresh operations will further damage the performance and lifetime of the flash memory. Therefore, we have proposed a precise refresh strategy to refresh only the data that needs to be refreshed, which reduces the wear of the flash memory device and improves the reading efficiency.

VI. ACKNOWLEDGEMENT

This work is supported by The National Natural Science Foundation of China under Grant No.61762075. It is also supported by The Provincial Nature Science Foundation of Qinghai under Grant No.2020-ZJ-926. Ping Xie is the corresponding author of this paper.

REFERENCES

[1] Cernea, R. A., Pham, L., Moogat, F., Chan, S., Le, B., Li, Y., ... & Quader, K. (2008). A 34 MB/s MLC write throughput 16 Gb NAND with all bit line architecture on 56 nm technology. IEEE Journal of Solid-State Circuits, 44(1), 186-194.

[2] Tokutomi, T., Tanakamaru, S., Iwasaki, T. O., & Takeuchi, K. (2014, May). Advanced error prediction LDPC for high-speed reliable TLC nand-based SSDs. In 2014 IEEE 6th International Memory Workshop (IMW) (pp. 1-4). IEEE.

[3] Zhao, K., Zhao, W., Sun, H., Zhang, X., Zheng, N., & Zhang, T. (2013). LDPC-in-SSD: Making advanced error correction codes work effectively in solid state drives. In 11th {USENIX} Conference on File and Storage Technologies ({FAST} 13) (pp. 243-256).

[4] Tanakamaru, S., Yanagihara, Y., & Takeuchi, K. (2013, January). Highly reliable solid-state drives (SSDs) with error-prediction LDPC (EP-LDPC) architecture and error-recovery scheme. In 2013 18th Asia and South Pacific Design Automation Conference (ASP-DAC) (pp. 83-84). IEEE.

[5] Liu, R. S., Chuang, M. Y., Yang, C. L., Li, C. H., Ho, K. C., & Li, H. P. (2014, June). EC-Cache: Exploiting error locality to optimize LDPC in NAND flash-based SSDs. In 2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC) (pp. 1-6). IEEE.

[6] Du, Y., Zou, D., Li, Q., Shi, L., Jin, H., & Xue, C. J. (2017, May). Laldpc: Latency-aware ldpc for read performance improvement of solid state drives. In Proc. MSST (pp. 1-11).

[7] Cai, Y., Yalcin, G., Mutlu, O., Haratsch, E. F., Cristal, A., Unsal, O. S., & Mai, K. (2012, September). Flash correct-and-refresh: Retention-aware error management for increased flash memory lifetime. In 2012 IEEE 30th International Conference on Computer Design (ICCD) (pp. 94-101). IEEE.

[8] Lv, Y., Shi, L., Li, Q., Gao, C., Xue, C. J., & Sha, E. (2019, August). Optimizing Tail Latency of LDPC based Flash Memory Storage Systems Via Smart Refresh. In 2019 IEEE International Conference on Networking, Architecture and Storage (NAS) (pp. 1-8). IEEE.

[9] Ryan, W. E. (2004). An introduction to LDPC codes. CRC Handbook for Coding and Signal Processing for Recording Systems, 1-23.

[10] Van Houdt, B. (2013). Performance of garbage collection algorithms for flash-based solid state drives with hot/cold data. Performance Evaluation, 70(10), 692-703.

[11] Zhou, B., Wan, S., & Xie, C. (2021). Isolation: Inexpensively separating cold data via garbage collection to improve the lifetime and performance of NAND flash SSDs. Concurrency and Computation: Practice and Experience, 33(15), e5460.

[12] Hwang, S. H., Kwak, J. W., & Park, C. H. (2015). Cold Data Identification using Raw Bit Error Rate in Wear Leveling for NAND Flash Memory. Journal of the Korea Society of Computer and Information, 20(12), 1-8.

[13] Luo, Y., Cai, Y., Ghose, S., Choi, J., & Mutlu, O. (2015, May). WARM: Improving NAND flash memory lifetime with write-hotness aware retention management. In 2015 31st Symposium on Mass Storage Systems and Technologies (MSST) (pp. 1-14). IEEE.

[14] Bucy, J. S., & Ganger, G. R. (2003). The DiskSim simulation environment version 3.0 reference manual. School of Computer Science, Carnegie Mellon University.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

A Power-on Hardware Self-test Framework in Web-based OS

Hao Xu*, Long Peng*, Jun Ma, Shasha Li, Jie Yu+, Qingbo Wu

College of Computer, National University of Defense Technology, 410073, China

***Abstract*—The hardware status of mobile devices is crucial to users. If hardware fails and running application cannot perceive it, it may cause application to process wrong data and eventually make the system fall into a logically chaotic state, which may lead to severe consequences. Web-based Operating System (OS) is a mobile OS on which the application is developed by Web, but there is no mechanism for users to get hardware status in the OS. To address the above issue, we propose a hardware power-on self-test framework to deliver hardware status of Web-based OS to users. We design the framework by deep stacking three layers from bottom to top: (i) infrastructure layer, in which we register a new driver named self-test to collect the information about the hardware status from other drivers. (ii) web engine layer, in which we implement a Web API named CallCLibrary to read the status from infrastructure layer and pass it to web application; (iii) web application layer, which receives status information from lower layer and displays it to users. We implement the framework on a device equipped with Firefox OS, which is representative of Web-based OS and conduct experiments to verify the self-test framework and test the performance of CallCLibrary. Experimental results show that the proposed framework successfully captures various failure statuses of hardware and CallCLibrary outperforms native Web APIs of Web-based OS and JNI of Android.**

***Keywords*-Web-based OS; hardware self-test; device drivers; Web API;**

I. INTRODUCTION

The mobile devices are widespread nowadays, which puts forward an urgent requirements for mobile operating system (OS). The Web-based OS is a type of mobile OS and all applications in this OS are developed on Web technologies. Prevalent Web-based OSs include Firefox OS [1] [2] , Chrome OS [3] and Kai OS. Although Firefox OS has been stopped by Mozilla in 2016, it has not disappeared in the market nowadays. Kai OS, based on Firefox OS, is currently very popular in India. Data reports show that as early as the first quarter of 2018, Kai OS surpassed iOS to become the second largest mobile operating system in India, second only to Android. In 2019, the number of users of Kai OS has exceeded 100 million, making it the third largest operating system in the world. Web-based OSs are generally composed of three layers, i.e. application, web engine and infrastructure layers [4], as shown in Fig. 1.

- **Application layer**, which consists of web applications that are written in HTML5, CSS and JavaScript, is the user interface layer and provides human-machine interactions. This layer is also responsible for interacting with lower layers through exposed APIs.
- **Web engine layer** is composed of Web APIs [5], layout engine, JavaScript engine and security module. Web APIs provide the programmatic access to hardware functionality and expose it to web application. Layout engine parses the HTML and renders the content to the screen. JavaScript engine specializes in processing JavaScript scripts.
- **Infrastructure layer** is the bottom of Web-based OS, which serves as a bridge between web engine layer and the underlying hardware. This layer controls the underlying hardware and exposes hardware capabilities to Web APIs that are implemented in web engine layer. This layer contains Linux kernel, device drivers and HAL (hardware abstract layer).

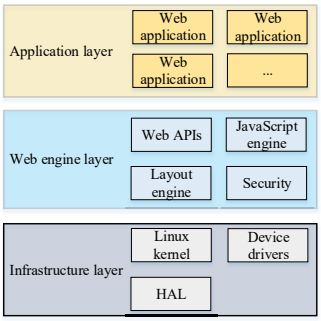


Figure 1: The architecture of Web-based OS.

System stability is an crucial indicator for an operating system, and whether the hardware works well or not is an important aspect of it. For instance, if GPS fails and device user fails perceiving it, the device may be in the wrong operating state and cause serious consequences, such as a map navigation system. At present, as far as we know, prevalent Web-based OS fails to provide hardware status information to users in the boot stage. To address above issue, we propose a power-on hardware self-test framework in Web-based OS, which can deliver hardware status to users. We implement the framework in three levels:

(i) In infrastructure layer we register a new driver named self-test to collect the status of other hardware drivers in the OS. To build the connection between the self-test driver and other hardware drivers, we utilize the notify mechanism in the Linux kernel. After the information collection, the self-test driver writes the hardware status into a file in the device.

* indicates equal contribution.

+ corresponding author. E-mail address: yj@nudt.edu.cn

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00061

(ii) In web engine layer we implement a Web API named CallCLibrary. CallCLibrary provides the web application (HTML5) with the ability to invoke native code such as C/C++. We use CallCLibrary to read the hardware status file and deliver the message to the web application. CallCLibrary is implemented complying the rule of Web API and utilizes the callback Web API to achieve the message delivery.

(iii) In application layer, we create a web application and receive the message delivered by CallCLibrary Web API, then display the message on the login interface in the boot stage.

To validate the usability of our framework and the performance of CallCLibrary, we conduct experiments on a same smart watch device equipped with Firefox OS and Android respectively. Firstly, we simulate hardware failure and experimental results show that the proposed framework could successfully capture it and deliver it to users. Then we validate the performance of CallCLibrary. The exiting mechanisms for the application invoking the native .so library include native Web API of Web-based OS and JNI (Java Native Interface) [6]of Android, and we conduct a comparison experiment between them. The experimental results show that CallCLibrary outperforms the native Web API and JNI, with the average invocation time 2.7ms compared to 3.88ms and 3.38ms.

II. Background

In this section, we present background knowledge to be used in the following sections. Firstly, we introduce the notify mechanism in Linux kernel which is responsible for the message delivery between different drivers in our framework. Then we describe the content related to hardware driver in Linux kernel. Finally, we introduce Web API which construct the connection between the web engine layer and the application layer.

A. Notify mechanism

The various subsystems in Linux are independent, but sometimes a subsystem may be of interest to other subsystems. To build the connection between different subsystems, the Linux kernel present s the notify mechanism.

The notify mechanism is implemented by a linked list in the Linux kernel. Each node in the notification list is a data structure named `notifier_block`, a `notifier_block` contains a function that is to be executed when the `notifier_block` is matched. For instance, subsystemA uses the function `atomic_notifier_chain _register` to register a `notifier_block` in the notify list, when an event happens in subsystemB and subsystemB wants to notify subsystemA, the subsystemB invokes the `notify_call_chain` to traverse all the `notifier_block` in notify list and find the registered `notifier_block`, then invoke the function in it when matched.

B. Driver management in Linux kernel

The driver is a bridge between hardware and operating system [7]. Since Linux 2.6, a set of driver management mechanism has been introduced and most drivers in Linux use this mechanism. Device in Linux is represented as `platform_device` and driver is represented as `platform_driver`.

There is a Linux kernel at the bottom of Web-based OS. When the device equipped with Web-based OS power on, the hardware device and drivers in the system start to register. Driver would look for the device with the same name in bus and hardware device does the same. The `probe` function in the driver will complete the final job when driver matches device successfully. And the `probe` function would check the registration about the hardware device, when the hardware device fails, the function would print error information in the kernel.

While the `probe` function test hardware in kernel mode which the user program cannot attach, resulting in users cannot get the status information about the hardware. Besides, the `probe` function test a single hardware driver and the information about all drivers need to be collected.

C. Web API

Web API is an application programming interface for Web, which could manipulate the DOM (Document Object Model) using JavaScript. The DOM is a document model in the browser and represent the document as a node tree, each node in the tree represents an element. Web API generally utilizes JavaScript object as a carrier. For instance, the Web API `document.getElementById(id)`, which could get the element corresponding to the id, depending on the `document` JavaScript object.

III. Power-on self-test framework

A. Overview

In the boot stage of the device equipped with Web-based OS, the `probe` function is invoked to check the respective driver registration in the Linux kernel. To notify users about the hardware driver status, there are two works to be done: i) Collect the device driver status information from all the hardware drivers. ii) Deliver the status information to users.

However, it is not easy to accomplish these works. The device driver program runs in respective subsystem, and there is no information exchange between various drivers. On the other hand, device driver program runs in kernel mode, while the web application program, which is the interface between users and Web-based OS, runs in user mode. The bridge between web application and device drivers is to be built.

To complete the above two works, we implement our framework in three levels: infrastructure layer, web engine layer (CallCLibrary Web API) and web application layer, as shown in Fig. 2. Next, we will give detailed descriptions of the three components in the framework.

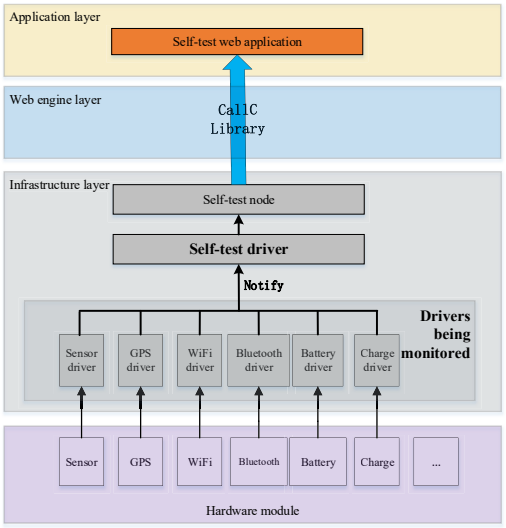


Figure 2: The architecture of the self-test framework.

B. Framework in Infrastructure layer

We register a new driver named self-test in the Linux kernel to collect information from other hardware divers through the notify mechanism and create a file to store the device status information. As a result, the information about hardware status is aggregated and the program in user mode could get the information from the device status file.

1) Register the self-test driver: The function `module_init` and `module_exit` exist in all the drivers of the Linux. `module_init` is the entry of the device driver, and when the device corresponding to the driver is deleted, the `module_exit` function will be executed. We create an entry function of the self-test driver and invoke the `subsys_initcall` instead of `module_init` to register the entry function, because the self-test driver needs to be registered before drivers to be monitored. The entry function contains function `platform _register_driver` which is used to register the self-test driver to bus.

2) Connect with other drivers through notify mechanism: The self-test driver is the crucial driver in the framework which is responsible for the information distribution and collection. On the other hand, the driver to be monitored delivers the hardware status message to the self-test driver. We implement the connection through `notify` mechanism in two aspects: the self-test driver and the driver being monitored .

Self-test driver. We use the `ATOMIC_NOTIFIER_HEAD` to initialize the notify list. Then we implement a `notifer_block` that contains the function `check_notifier_call`. The function `check_notifier _call` identifies the information delivered from drivers being monitored and set the device status file. Next we complete the implementation in the `probe` function of the self-test driver. We invoke function `atomic_ notifier_chain_register` to register the `notifer_block` mentioned above into the notify list.

Drivers being monitored. We encapsulate the function `notify_call_chain` into a function `notify_check_notifier` and add the function to the proper location of the `probe` function in the driver code. The function `notify_check_notifier` delivers "SUCCESS" or "FAIL" message to the notify list. For instance, in the `probe` function of compass driver file, we add the function `notify_check_notifier` as shown in Fig. 3.

```
kernel/linux-3.10.y/drivers/comip/sensor/akm09911.c
1467.int akm_compass_probe(struct i2c_client *client, const struct i2c_device_id *id)
1468.{
                        …
1479.   if (!pdata) {
1480.           dev_err(&client->dev, "Failed to allocate memory\n");
1481.           printk("9999===akm_compass_probe allocate memory failed \n");
1482.           notify_check_notifier(CHECK_NOTIFY_COMPASS_FAIL);
1483.           return -ENOMEM;
1484.   }
                        …
1486.   err = akm_parse_dt(&client->dev, pdata);
1487.   if (err) {
1488.           dev_err(&client->dev, "DT parsing failed\n");
1489.           printk("9999===akm_compass_probe DT parsing failed \n");
1490.           notify_check_notifier(CHECK_NOTIFY_COMPASS_FAIL);
1491.           return err;
1492.   }
                        …
1635.           dev_info(&client->dev, "successfully probed.");
1636.           notify_check_notifier(CHECK_NOTIFY_COMPASS_SUCCESS);
1637.                   return 0;
```

Figure 3: The code snippet added in the sensor driver.

When the `notify_call_chain` in the `notify_check_notifier` is invoked, the notify list will find the `notify_block` registered before and execute the function (i.e. `check_notifier_call`) in it to update the hardware status file.

C. Framework in web engine layer (CallCLibrary Web API)

In Web-based OS, web application is the interface between users and operating system, so the information in hardware status file need to be delivered to web application. We propose the CallCLibrary Web API to achieve that goal.

In this section, firstly we introduce the CallCLibrary Web API, which provide the ability to invoke the native .so library for web application. Next we present the specific application of the CallCLibrary on self-test framework.

CallCLibrary is a type of Web API, so we construct it complying the rule of Web API. The overall design of the CallCLibrary Web API is shown in Fig. 4. The implementation is divided into three parts: application layer, web engine layer and plugin (i.e. .so library).

1) CallCLibrary in application layer: In application layer, i.e., the web application, we construct a CallCLibrary JavaScript object. Three functions of CallCLibrary are `Init`, `Addlistener` and `Exec`. `Init` is used to

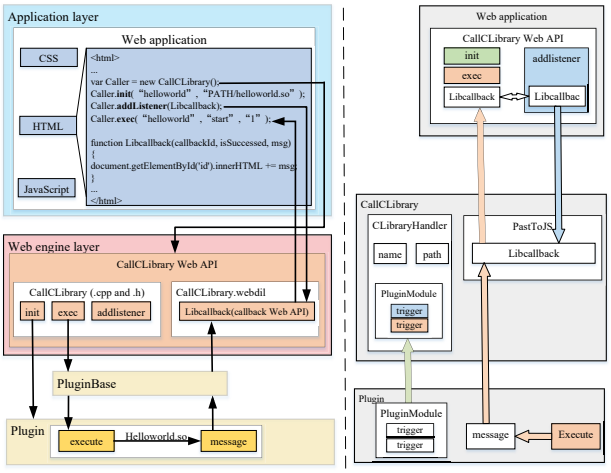


Figure 4: The architecture of CallCLibrary Web API. The left is the overview and the right is the specific implementation with data structure.

identify the corresponding .so library and complete the work for initialization; `Addlistener` adds a callback function to the web engine layer, the message about the .so library would be delivered to the web application through the callback function; `Exec` notifies the .so library to perform the specific actions and trigger the callback function added by `Addlistener` to deliver the message to web application. The parameters of `Exec` include name of the .so library ("`helloworld`"), the specific action about execution ("`start`") and the callbackid ("`1`"), as shown in Fig. 4. The three parameters of the `Libcallback` are callback id ("`callbackid`"), whether the callback function executes successfully ("`isSuccessed`"), and the delivered message ("`msg`"). All the parameters of `Libcallback` are delivered from web engine layer through the callback Web API.

2) CallCLibrary in web engine layer: The specific implementation of `Init`, `Addlistener`, `Exec` is in web engine layer and it is the major part of CallCLibrary. We construct the CallCLibrary Web API through the `CallCLibrary.webidl` and the implementation files, which build the connection between application layer and web engine layer . Then we create a class named `PluginBase` to connect the plugin .so library with web engine layer. In our CallLibrary framework, the plugin .so library only has one class, and the class need inherit the class `PluginBase`, the specific operation about the plugin class need add to the function `execute`. The work about connection with web engine layer is done by `PluginBase`, the only work for the plugin .so library is fill the function `execute`.

Similar with the regular Web API, we create the moz.build file and `CallCLibrary.webidl` as shown in Fig. 5. We define the three functions of the CallCLibrary and declare a callback Web API named `LibCallback`. The callback Web API is used to deliver the callback function in Web-based OS and the callback information would be delivered to the parameters of the function `Libcallback` in web application.

```
1 [Constructor]
2 interface CallCLibrary{
3   boolean init(DOMString lib, DOMString path);
4
5   DOMString exec(DOMString lib, DOMString function,
6           DOMString callbackid);
7
8   void addListener(LibCallback libCallback);
9 };
10
11  callback LibCallback = void (DOMString callbackId,
12          boolean success,DOMString msg);
```

Figure 5: The code snippet of CallCLibrary.webidl.

Next we will present the specific implementation about the CallClibrary. The related data structure in implementation is shown in TABLE. I. We show the implementation in three phases: Initialization of CallCLibrary, adding callback function and function execution in .so library through CallCLibrary.

Initialization of CallCLibrary. In this phase, firstly we get the name and the path of the .so library from web application through webidl. Next we use the function `dlopen` and `dlsym` to get the `PluginModule` in the `PluginBase` and store the `PluginModule` into the `CLibraryhandler`.

Adding callback function. The `Libcallback` from web application will be delivered to web engine layer through the callback Web API, then the trigger function in `PluginModule` adds the `Libcallback` to `PastToJS`.

Execution. The trigger function in `PluginModule` of the ClibraryHandler will notify the .so library to invoke the function `execute`, and execute the specific action in plugin which is implemented before. After the execution, the message to be delivered to web application is generated. The plugin notifies the `PastToJS` to get the `Libcallback` and invokes the `call` API of the `Libcallback` with the message as a parameter in it, then the message about the C/C++ plugin execution would be delivered to the web application through the `LibCallback`.

3) CallCLibrary in plugin (i.e., .so library): All the C/C++ plugins must inherit the class `PluginBase` before being compiled to .so library. The only work for the plugin is completing the function `execute`. The `PluginBase` is the bridge between the plugin and the CallCLibrary. Until now, we describe the specific implementation about CallClibrary Web API, then we will introduce the implementation on our self-test framework.

4) CallCLibrary on our self-test framework: We create a plugin named self-test complying the rule about CallCLibrary and compile it to `self-test.so`. In the self-test plugin, we use the function `fopen` and `fread` to get the message about the device status file. Then we deliver the message to self-test web application by CallCLibrary Web API.

Table I: The Data structure in the implementation about CallLibrary.

Data structures	Description
PluginModule	It exist in PluginBase and CLibraryHandler. The specific implementation is in PluginBase and it would be delivered to CLibraryHandler in the initialization. It contains trigger functions for adding callback function and contains executing function in web engine layer.
CLibraryHandler	The most important data structure in the CallCLibrary. It contains the name and the path of the .so library, and stores a PluginModule which is delivered from PluginBase in the initialization.
PluginBase	The superclass of all the plugins (.so library) in CallCLibrary. It contains a PluginModule and several utility functions, which are responsible for the connection with web engine layer.
PastToJS	In the adding callback function stage, it stores the Libcallback delivered from web application; In the execution stage, it invokes the Libcallback Web API and delivers the callback message to web application.

D. Framework in web application layer.

We develop a web application named self-test which is written in HTML5 and put it in the boot stage of the Web-based OS. The web application receives the hardware status information through the CallCLibrary Web API and displays the information to users.

IV. Evaluations

In this section, we present the experimental study of our proposed self-test framework. Firstly we introduce the setup of the experiments. Next, we conduct two experiments: (i) Verify whether the self-test framework could identify the hardware failure. (ii) Verify the performance of CallCLibrary Web API.

We run the experiments on a smart watch device equipped with Web-based OS (Firefox OS version 44.0) and Android (version 6.0.0) respectively, with its CPU 1.6 GHz, 4 GB RAM. We choose the ADB (Android Debug Bridge) [8] to control the device on a Ubuntu 18.04 desktop platform containing Intel (R) Core (TM) i7 CPU 1.8 GHz with 16 GB RAM.

A. Self-test framework verification.

1) Experiment : The hardware in the smart watch include sensor, Wi-Fi, GPS, Bluetooth, battery and so on. The hardware malfunction is not easy to be created artificially and we conduct the simulation. Take the GPS for example, we change its access permission to deter the web application from accessing it, then the GPS fails to work. We conduct the experiment on the same smart watch device with simulation and no simulation respectively, and observe the self-test framework whether could detect the failure.

2) Results : We press the power-on key, self-test framework shows the GPS works well. Then we conduct the failure simulation and the self-test framework detects the failure as shown in Fig. 6.

B. CallCLibrary performance

We conduct an experiment on the performance of CallCLibrary. The experiment compares CallCLibrary with native Web API and JNI in Android.

Figure 6: The self-test result about GPS.

1) Preliminaries: All three APIs mentioned above could invoke the function in .so library from application level. The CallCLibrary and native Web API are in Web-based OS on which the application is written in HTML5; the JNI is in Android on which the application is written in Java.

Native Web API. The Web-based OS provides Web API which could manipulate the hardware. The web engine layer handles the low-level access to the system using a C++ API that is accessible to the higher levels. For instance, when users need the device vibrate, the web application in application layer submits requests to access to underlying device via vibrate Web API, then the web engine layer submits the request to the infrastructure layer. The single request in web application layer would result in operations in infrastructure layer. In HAL of the infrastructure layer, the .so library implements the connection between the hardware driver and the web engine. The whole process is shown in Fig. 7. Compared with native Web API, the CallCLibrary Web API provides a more flexible interface for higher levels of the Web-based OS, and we could add what we want in the .so library as long as the rule of the CallCLibrary is complied. If we want to add the interface for a new hardware, we could use CallCLibrary instead of adding a new single implementation like the vibrate Web API, in other words, the CallClibrary is more general.

Android JNI. The JNI achieve the inter operation between the Java and native libraries. JNI technology enables Java program to run on JVM (Java virtual machine) [9], pass the parameters to native codes such as C/C++ and get the return values. The invocation in Java attach the .so library in HAL through JNI.

Web-based OS and Android. When developer design

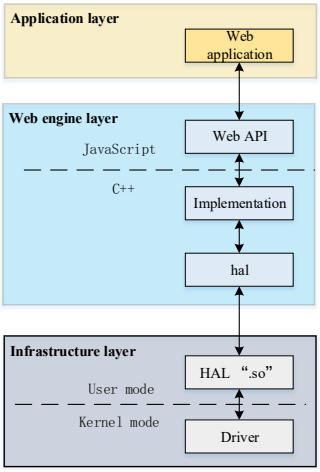


Figure 7: The native Web API in Web-based OS.

the Web-based OS, Android is well-established in the hardware-adaption, so they apply this part to the Web-based OS (i.e., Infrastructure layer) , basing on the AOSP (Android Open Source Project) [10]. Above the infrastructure layer, the Web-based OS transplants browser engine to support the HTML5 application. By contrast, Android use the JVM to support the Java application. In a word, Web-based OS and Android are very similar at the bottom layer such as device driver, which is the basis of experiment B.

2) Experimental design: In this experiment, we test three APIs, which could invoke native function in application: CallCLibrary Web API invokes the function `execute` in `self-test.so`, Vibrate Web API and Vibrate interface in Java (JNI) invoke the function `vibrate_on` in `libhardware_legacy.so`. All the .so library are under the `/system/lib/` directory of the device.

In Web-based OS, we develop a web application to test CallCLibrary and Vibrate Web API. We add a timestamp at the function `exec` of the CallCLibrary which is the entry of the invocation, then we add another timestamp to the function `execute` of the self-test.c, which is to be compiled to `self-test.so`; The Vibrate Web API could make the device equipped with Web-based OS Vibrate. We add `window.navigator.Vibrate()` to the web application, which is the invocation entry and we add a timestamp here, then we add another timestamp to the function `vibrate_on` of the vibrate.c, which is to be compiled to `libhardware_legacy.so`. The `libhardware_legacy.so` belongs to the HAL of the Web-based OS.

In Android, we test the vibrate interface in Java. We develop a vibrate android application with Android Studio (version 4.2.2) In Java application, we import the package `android.os.Vibrator` and get the object vibrator with the function `getSystemService (Service.VIBRATOR_SERVICE)`. Then we invoke the function `vibrator.vibrate(1000)` which means vibrate 1000ms on the device and add a timestamp here. Next, similar with the Vibrate Web API, we add another timestamp to the function `vibrate_on` in vibrate.c.

We conduct the invocation 50 times on each APIs (CallCLibrary Web API, native Web API, JNI). The invocation time $\triangle T = T2 - T1$, T2 is the time when the function in .so library being invoked and T1 is the time when the API in application(Web or Java) being invoked. We utilize the ADB tools to observe the experiment result and input the "adb logcat" in the terminal of PC and get the according time.

3) Results and discussions: As shown in Fig. 8, the CallCLibrary outperforms native Web API and JNI. The invocation time about CallCLibrary is smoother than others. The average time in 50 invocations for CallCLibrary is 2.7ms, 3.88ms for native Web API and 3.38ms for JNI.

The CallCLibrary directly invokes the .so library while the native Web API has to complete the work in HAL before invokes the .so library, so is the JNI, which consume extra time. We propose the CallCLibrary to provide a more flexible method for application to invoke the function in .so library, and the performance is acceptable.

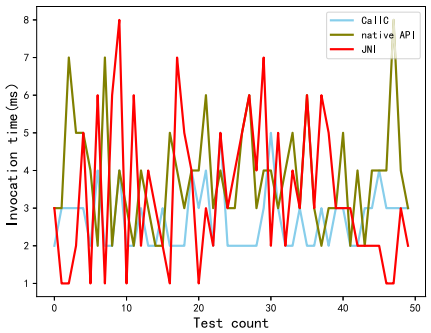


Figure 8: The invocation time about CallCLibrary, native Web API and JNI.

V. RELATED WORKS

A. Fault detection on mobile OS

We mainly focus on Android operating system, because the bottom of Web-based OS is based on AOSP. The relationship between Android and Web-based OS is introduced in Experiment B. Android fault detection consists of hardware and software detection. The software detection consists of two approaches, namely, static and dynamic analysis [11]. As for the hardware detection, the study is mainly focus on the driver analysis [12]–[14], the power-on hardware test using the driver registration is not involved.

B. C/C++ codes on Web

The CallCLibrary could invoke native code in web application which is developed in a Web environment. At present, there are two methods for Web to invoke native code such as C/C++.

Compile the C/C++ code to the language the browser could identify. The asm.js [15] is a subset of JavaScript

and is mainly used as compilation target. Developers start with an existing C/C++ application which they can then efficiently port to the web by compiling it to asm.js using Emscripten [16]. WebAssembly [17] is a new byte code designed for the web. It also provides a compilation target for languages such as C/C++. The Web Assembly cannot directly access Web APIs. The two methods mentioned above is used for run the native application which is compiled before on the web, as a comparison, we complete the invocation through CallClibrary Web API on web which is more flexible, we could invoke the native code whenever we want.

Invoke the C/C++ dynamic linked library in the browser. The NPAPI (Netscape Plugin Application Programming Interface) could make the native code run as part of the web application through invoking the C/C++ `.so` library. While the plugin developed by NPAPI runs at the same level as the browser, which is a huge security risk to the system, the NPAPI is blocked by mainstream browser since 2014. Node.js is a server-side JavaScript environment built on Google Chrome's v8 engine. There are two main ways for Node.js to use the function of the C/C++layer, the first is to call the global variables process, Buffer, etc., and the second is the function process.binding. Node.js does not apply to the scenario in this paper.

VI. CONCLUSIONS

The hardware status is a major security part of an mobile OS. Web-based OS is a type of mobile OS and there is not a mechanism for the hardware status test in the OS. So we proposed a self-test framework in the boot stage of the device equipped with Web-based OS. In the implementation of the self-test framework, we propose a Web API named CallCLibrary to facilitate the upper application invoke the native code. As a result, the hardware status information is delivered from the file in bottom of the OS to upper application through the CallCLibrary Web API. In experiments, we verify the self-test framework and the performance of CallCLibrary. The experiments show that the framework could identify the hardware failure accurately and CallCLibrary outperform other similar technologies such as native Web API and JNI.

REFERENCES

[1] M. Jadhav and K. K. Joshi, "Forensic investigation procedure for data acquisition and analysis of Firefox OS based mobile devices," 2016 International Conference on Computing, Analytics and Security Trends (CAST), 2016, pp. 456-461

[2] M. N. Yusoff, R. Mahmod, M. T. Abdullah and A. Dehghantanha,"Mobile forensic data acquisition in Firefox OS," 2014 Third International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec), 2014, pp. 27-31

[3] I. Bente, B. Hellmann, T. Rossow, J. Vieweg and J. von Helden, "On Remote Attestation for Google Chrome OS,"2012 15th International Conference on Network-Based Information Systems, 2012, pp. 376-383

[4] B2G architecture, 02/06/2021. [Online]. Available: https://developer.mozilla.org/enUS/docs/Archive/B2G_OS/Architecture

[5] Web API, 03/06/2021. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/API

[6] C. Qian, X. Luo, Y. Shao and A. T. S. Chan, "On Tracking Information Flows through JNI in Android Applications," 2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, 2014, pp. 180-191
2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), 2021, pp. 1708-1718

[7] I. Pustogarov, Q. Wu and D. Lie, "Ex-vivo dynamic analysis framework for Android device drivers," 2020 IEEE Symposium on Security and Privacy (SP), 2020, pp. 1088-1105

[8] ADB, [online] Available: https://developer.android.com/studio/command-line/adb

[9] Chien-Wei Chang, Chun-Yu Lin, Chung-Ta King, Yi-Fan Chung and Shau-Yin Tseng, "Implementation of JVM tool interface on Dalvik virtual machine," Proceedings of 2010 International Symposium on VLSI Design, Automation and Test, 2010, pp. 143-146

[10] X. Song and C. Yang, "Mobile Device Management System Based on AOSP and SELinux," 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), 2017, pp. 417-420

[11] L. Taheri, A. F. A. Kadir and A. H. Lashkari, "cExtensible Android Malware Detection and Family Classification Using Network-Flows and API-Calls," 2019 International Carnahan Conference on Security Technology (ICCST), 2019, pp. 1-8

[12] S. M. S. Talebi, H. Tavakoli, H. Zhang, Z. Zhang, A. A. Sani, and Z. Qian, "Charm: Facilitating dynamic analysis of device drivers of mobile systems," in 27th USENIX Security Symposium (USENIX Security 18). Baltimore, MD:USENIX Association, 2018, pp. 291–307

[13] D. Davidson, B. Moench, T. Ristenpart, and S. Jha, "FIE on firmware: Finding vulnerabilities in embedded systems using symbolic execution," in Presented as part of the 22nd USENIX Security Symposium (USENIX Security 13). Washington, D.C.: USENIX, 2013, pp. 463–478.

[14] D. Song, F. Hetzelt, D. Das, C. Spensky, Y. Na, S. Volckaert, G. Vigna, C. Kruegel, J.-P. Seifert, and M. Franz, "Periscope: An effectiveness probing and fuzzing framework for the hardware-OS boundary," in 2019 Network and Distributed Systems Security Symposium (NDSS). Internet Society, 2019, pp. 1–15.

[15] Noah Van Es, Jens Nicolay, Quentin Stievenart, Theo D'Hondt, and Coen De Roover. "A performant scheme interpreter in asm.js," in 31st Annual ACM Symposium on Applied Computing (SAC 16). New York, Association for Computing Machinery, 2016, pp.1944–1951.

[16] A. Zakai, "Fast Physics on the Web Using C++, JavaScript, and Emscripten," in Computing in Science & Engineering, vol. 20, no. 1, pp. 11-19, January/February 2018.

[17] A. Romano and W. Wang, "cWASim: Understanding WebAssembly Applications through Classification," 2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE), 2020, pp. 1321-1325.

Prestress Optimal Design of Deployable Antenna Considering The Effect of Gravity

Guanlong Su, Yang Li, Yesen Fan, Huanxiao Li, Xiaofei Ma*
Xi'an Institute of space radio technology
Xi'an, China
e-mail: 313104243@qq.com

***Abstract*—The prestress of deployable antenna reflector is designed for vacuum environment and tested on the ground. Which implies the effect of gravity has to be considered in order to get higher surface accuracy. In this paper, a new method for prestress design under gravity is proposed. A typical antenna mesh is simulated as an example. Results demonstrate the antenna mesh can be restored to the ideal shape after removing the gravity.**

Keywords-antenna reflector; prestress design; gravity effect;

I. INTRODUCTION

With the vigorous development of aerospace technology, the demand for large-scale space deployable antennas is becoming more and more urgent. In order to satisfy the requirements such as tiny signal transmission on the ground, greater capacity of information transmission, and the realization of high resolution of remote sensing, high accuracy of antenna mesh has become a key issue.

The mesh reflector of a deployable antenna is known as a cable net structure, which belongs to the family of flexible tension structures characterized by strong geometric nonlinearities. The initial stiffness and shape can be achieved by prestress design. The purpose is to find a surface sufficiently close to a desired profile under specific tension loads[1]. The prestress distribution plays a vital role on surface accuracy of mesh reflector antennas[2].

The prestress design of mesh reflector antennas is an extremely challenging problem resulting from the large scale and flexibility. However, some studies are put out only considering the cable-net[3,4]. An equal tension method is proposed to improve the stability of deployable antenna[5]. Two methods are proposed considering the coupling of supporting trusses and flexible cable nets in [6]. Shape control concepts for mesh reflectors have been explored by two ways as well [7,8].

However, these existing methods based on the force density method are only suitable in ideal environment. That means no gravity effect is taken into consideration. For a real antenna reflector, it should be adjusted on ground where gravity could not be ignored. Moreover, the influence for surface accuracy resulted by gravity increases as the size of the antenna increases. The main purpose of this paper is to propose a new method to optimize the prestress in the cable-net structure. From which a new model under gravity can be achieved to supply guidance in antenna adjustment. Simultaneously, this model also has the ability to recover to the desired surface in orbit.

II. PRESTRESS OPTIMAL DESIGN METHODS

A. Composition of a mesh reflector antenna

A deployed mesh reflector antenna is conceived with the concept of a tension truss, which is a light and inherently stiff structure that can be precisely and repeatedly deployed regardless of environment. As illustrated in Fig. 1, it is divided into three parts: a supporting truss, boundary cables, and a cable net reflector including surface cables and tension tie cables. The cable net reflector and the boundary cables together are named as the deployable cable net structure.

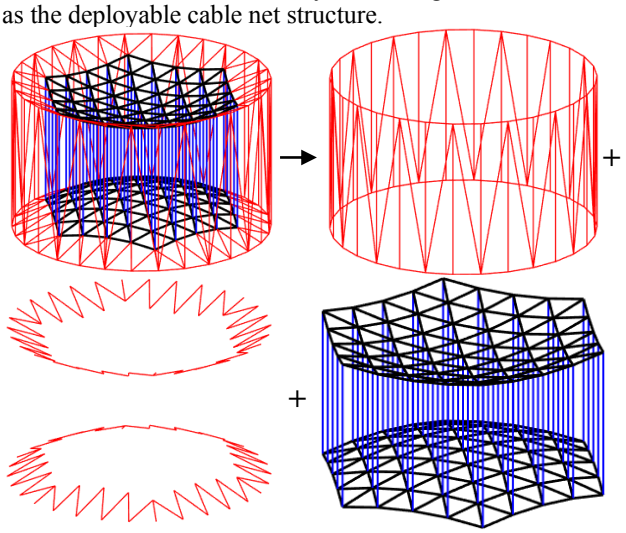


Figure 1. Composition of a mesh reflector antenna

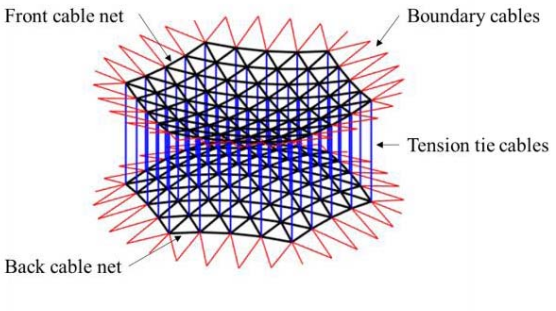


Figure 2. Schematic of a cable net structure

From the geometric viewpoint, a cable net structure can also be divided into: a front cable net, a back cable net, and tension tie cables, as shown in Fig. 2. Front and back cable nets are both doubly curved geodesic nets that are

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00062

placed back-to-back. Tension tie cables evenly apply approximately normal forces between front and back cable nets to permanently preload them in tension to maintain the surface profile of the reflector.

B. *Prestress design considering gravity effect*

The prestress design of a cable net structure aims to determine cable tension distribution to obtain the required reflector surface accuracy. A unit of a cable net structure is shown in Fig. 3, where node *i* is connected to node *j* by a cable. The equilibrium equation of node i can be derived as follows:

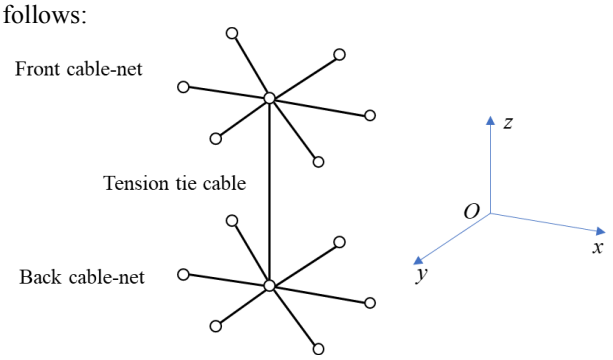


Figure 3. A cable-net unit

$$\begin{cases} \sum_j F_{ij}\dfrac{x_i - x_j}{l_{ij}} = 0 \\ \sum_j F_{ij}\dfrac{y_i - y_j}{l_{ij}} = 0 \\ \sum_j F_{ij}\dfrac{z_i - z_j}{l_{ij}} = \sum_j G_{ij} \end{cases} \tag{1}$$

where F_{ij} , G_{ij} and l_{ij} , respectively, represent the tension force, the gravity and the length of the cable between two adjacent nodes i and j of coordinates (x_i, y_i, z_i) and (x_j, y_j, z_j) . It can be rewritten in the form of matrix as follows[9]:

$$\begin{cases} \mathbf{C^T Q C x + C^T Q C_f x_f = 0} \\ \mathbf{C^T Q C y + C^T Q C_f y_f = 0} \\ \mathbf{C^T Q C z + C^T Q C_f z_f = G} \end{cases} \tag{2}$$

where $\boldsymbol{Q}$, $\boldsymbol{C}$ and $\boldsymbol{C}_f$, respectively, represent the force-density matrix, the topology matrix of free nodes and the topology matrix of boundary nodes. The boundary nodes are considered to be fixed.

With a given load and a given position of fixed points we get for each set of prescribed force densities exactly one equilibrium state with the shape

$$\begin{cases} \mathbf{x = -(C^T Q C)^{-1} C^T Q C_f x_f} \\ \mathbf{y = -(C^T Q C)^{-1} C^T Q C_f y_f} \\ \mathbf{z = (C^T Q C)^{-1} (G - C^T Q C_f z_f)} \end{cases} \tag{3}$$

Since the assumption of nodes located in the required parabolic surface, the nodal locations and topology of the cable net structure can be determined. Then the prestress design of the cable net structure can be expressed by the following optimization model:

$$\begin{cases} Find \quad Q = diag[q_1, q_2, \cdots, q_n] \\ Min \quad RMS = \sqrt{(\sum_{j=1}^{n} \Delta d_j^2)/n} \\ s.t. \quad C^T QCD + C^T QC_f D_f = b \end{cases} \tag{4}$$

where

$$D = \begin{bmatrix} \mathbf{x} & & \\ & \mathbf{y} & \\ & & \mathbf{z} \end{bmatrix} \quad D_f = \begin{bmatrix} \mathbf{x}_f & & \\ & \mathbf{y}_f & \\ & & \mathbf{z}_f \end{bmatrix} \quad b = \begin{bmatrix} 0 \\ 0 \\ G \end{bmatrix}$$

Δd_j is the displacement of the jth free node in the cable net reflector, n is the total number of free nodes in the cable net reflector and RMS is an abbreviation of root mean square.

It should be noticed that the length of cable changes in each iteration step so that the gravity matrix also changes in each step. However, the original lengths of all cables remain unchanged during the optimization process.

C. *Prestress optimization*

For a given antenna surface, the prestress can be designed through "(4)" with the influence of gravity at the same time. Nevertheless, the position of the antenna reflector is unknown. Specifically, when gravity is not taken into consideration, the design surface is simply the ideal surface P0 in orbit. But the design surface under gravity would deformed to P2, which means that the target surface is unknown during the optimization. The difference can be illustrated in Fig. 4. As gravity only work in z-direction, the influence of the coordinates in the x and y directions are ignored. There is a deviation Δz ,which is unknown, between the target shape under gravity and the ideal shape. But for every certain value of Δz_i , prestress can be designed by "(4)" for the ith target shape. Moreover, this under-gravity surface should have the ability to restore to ideal position in orbit as required. Gravity is removed after the design so that the shape would deform to a new position P2 as shown in Fig. 4. The smaller the gap between P0 and P2, the better it is. Obviously, all values of Δz for free nodes are negative and for boundary nodes the values are 0. The solution is to find a certain Δz with which prestress design can be solved. Then the surface is able to recover to ideal surface after removing the gravity. The overall algorithm is presented in the flowchart of Fig. 5.

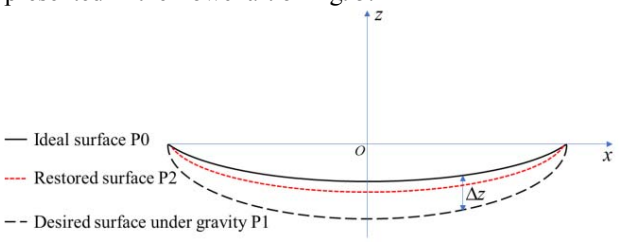


Figure 4. Schematic diagram of the relationship between the surfaces

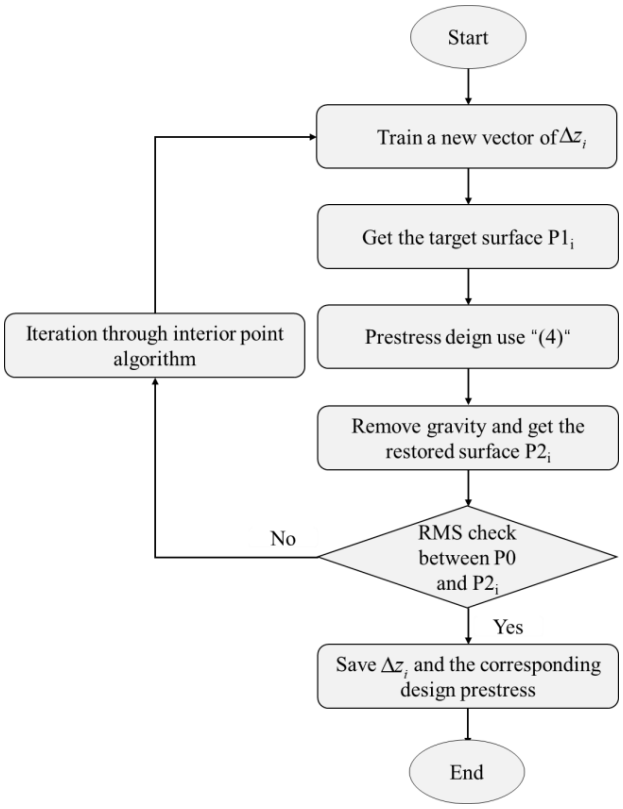


Figure 5. Algorithm of prestress optimization

The selection of the initial values of the iteration is a critical factor in the optimization. An improper selection of the initial value will cause the result to be a local optimum instead of a global optimum. And the iteration may not converge in worse cases.

III. NUMERICAL SIMULATION

A parabolic antenna with a spatial mesh reflector is shown in Fig. 1. The antenna specifications are as follows:

Diameter of aperture: 10,000 mm
Focal length of front cable net: 7,500 mm
Focal length of back cable net: 7,500 mm
Number of surface cables: 312 (=156×2)
Number of boundary cables: 108 (=54× 2)
Number of tension tie cables :61
Number of free nodes: 60 (=30×2)
Height: 2,000 mm
Type of facets: triangular
Elastic modulus of cables: 20 GPa
Radius of cables: 0.5mm
Density of cables: 1450kg/m^3

First, it is important to set the initial value of Δz, which can be expressed as

$$\Delta z = [\Delta z_1, \Delta z_2, \cdots, \Delta z_{n_f}, \underbrace{0, 0, \cdots 0}_{n_b}] \qquad (5)$$

Where n_f and n_b represents the number of free nodes and boundary nodes respectively. With the help of finite element analysis in ANSYS, the minimum value of Δz is set to be -4mm. And all values of Δz is constrained to be negative. Second, the tension forces **F** of all cable are set to 10N initially. The iteration results are shown in Fig.6. It can be seen that the RMS error of front cable net is lowered from 11.07 mm to 0.66 mm.

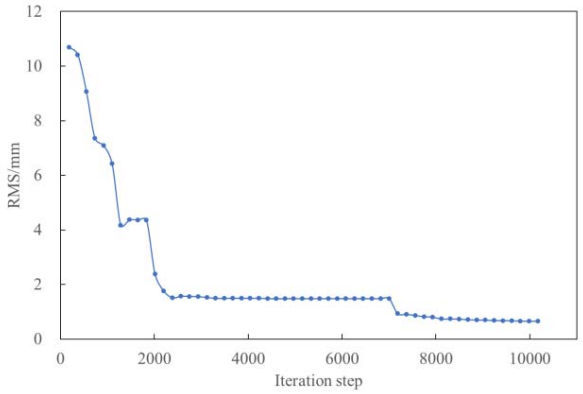


Figure 6. Iteration process of RMS error of the front cable ne

The deformation under gravity in z-direction is shown in Fig.7. The maximum nodal displacement is 3.79 mm. The prestress distribution of the cable net structure is shown in Table 1.

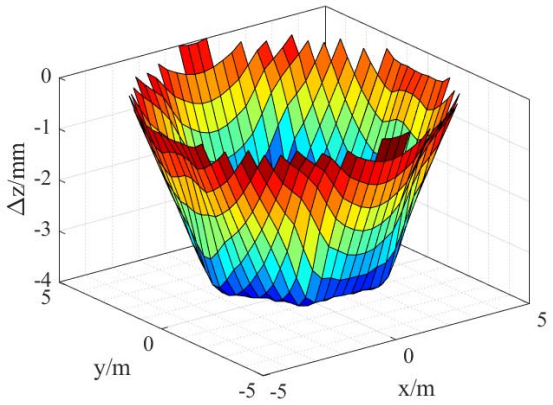


Figure 7. Deformation under gravity

TABLE I. PRESTRESS DISTRIBUTION OF THE CABLE NET STRUCTURE

Item	Prestress values in cables (N)		
	Maximum	***Minimum***	***Mean***
Front cable net	12.15	11.14	11.44
Back cable net	12.18	11.12	11.44
Tension tie cables	2.65	2.23	2.37

As shown in Fig.7, the distribution of Δz basically in line with the prediction of surface. Which confirms the validity of prestress optimization. On the other hand, the forces in Table.I show excellent uniformity that indicates a good stability in design result. The maximum and minimum stress ratios of the front cable net and the back cable net are 1.062 and 1.096 respectively. More importantly, the surface we find have the ability to restore to ideal surface when removing gravity. The RMS between two surfaces is 0.66mm which is acceptable in engineering. However, the algorithm iterated more than 10,000 times so

that more efficient method should be explored in future work.

IV. CONCLUSION

A new method for prestress design of deployable antenna reflectors has been proposed. Because of all the adjustments for deployable antennas are executed on ground, the influence of gravity becomes a key factor affecting the antenna surface accuracy, and this is particularly evident when the antenna size increases.

A parabolic antenna is simulated to validate the method. Results show the surface we found can well return to the ideal surface after removing the gravity. The prestress design shows good stability as well.

REFERENCES

[1] D. Yang, Y. Zhang, P. Li, and J. Du, "Numerical form-finding method for large mesh reflectors with elastic rim trusses," Acta Astronautica, vol. 147, 2018, pp.241-250.

[2] T. Li, "Deployment analysis and control of deployable space antenna," Aerospace Science and Technology, vol. 18(1), 2012, pp.42-47.

[3] G. Tibert, "Optimal design of tension truss antennas," 44th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference, 2003, pp.1629

[4] B. Yang, H. Shi, M. Thomson and H. Fang, "Optimal design of initial surface profile of deployable mesh reflectors via static modeling and quadratic programming," 50th AIAA/ASME/ASCE/AHS/ ASC structures, structural dynamics and materials conference, 2009, pp.2173.

[5] Y. Fan, X. Ma, Z. Li and T. Li, "Analysis Method of Influence of Cable Length Errors on Surface Accuracy of Cable-Net Antennas," Journal of South China University of Technology, vol. 8, 2014, pp.64-69.

[6] T. Li, J. Jiang and H. Deng, "Form-finding methods for deployable mesh reflector antennas." Chinese Journal of Aeronautics, vol. 26(5), 2013, pp.1276-1282.

[7] A. Meguro, S. Harada and M. Watanabe, "Key technologies for highaccuracy large mesh antenna reflectors," Acta Astronaut, vol. 53(11), 2003, pp.899-908.

[8] H. Tanaka, N. Shimozono and M. Natori, "A design method for cablenetwork structures considering the flexibility of supporting structures," Transactions of the Japan Society for Aeronautical and Space Sciences, vol. 50(170), 2008, pp.267-273.

[9] H. Schek, "The force density method for form finding and computation of general networks." Computer Methods in Applied Mechanics and Engineering, vol. 3(1), 1974, pp.115-134.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Electrostatic interaction investigations of the Barnase-Barstar complex using molecular dynamics simulation

Yi Fu[1,2]
1. School of Internet of Things Engineering,
Wuxi City College of Vocational Technology
Wuxi, China
2. Wuxi Research Center for Environmental Science & Engineering
fuyi@wxcu.edu.cn

Ji Zhao [1,2]
1. School of Internet of Things Engineering,
Wuxi City College of Vocational Technology
Wuxi, China
2. Wuxi Research Center for Environmental Science & Engineering
ji_zhaowx@163.com

Juan Mei[1,2]
1. School of Internet of Things Engineering,
Wuxi City College of Vocational Technology
Wuxi, China
2. Wuxi Research Center for Environmental Science & Engineering
meijuan_wx@163.com

***Abstract*—Protein-protein interactions can regulate many biological processes. Studying the mechanisms of protein-protein interactions is critical for understanding the biological functions and regulation mechanisms of proteins. The Barnase-Barstar complex is a well-studied target for computational analyses due to its simplicity and small size. Despite extensive studies, it is unclear whether there are differential responses to thermal stress between the Barnase-Barstar complex and the Barnase alone. In this study, molecular dynamics simulations were used to study intermolecular interactions and structural changes of the Barnase-Barstar complex under thermal stress. We performed molecular dynamics simulations under different temperatures for the complex and for Barnase itself. Our simulations suggest that under thermal stress, Barnase is more stable within the Barnase-Barstar complex than the protein itself, and the higher stability is likely attributed to the binding interactions between the two proteins. These results provide a theoretical foundation and an important guidance to understand mechanisms of biological regulation of protein complexes.**

Keywords- Barnase-Barstar complex; salt bridge; electrostatic interaction; molecular dynamics simulation

I. INTRODUCTION

Protein-protein interactions play an important role in biological processes, such as signal transduction, cell regulation, and the immune response [1]. Studies of these interactions are important for understanding biological functions and regulation mechanisms [2-3]. Among protein-protein interactions, binding of different proteins to form protein complex is widespread in all organisms. In *Bacillus amyloliquefaciens*, the Barnase-Barstar complex is a well-studied system, partly due to its simplicity and fast association rates [4]. Barnase is an extracellular ribonuclease with 110 residues from *Bacillus amyloliquefaciens*, and Barstar is a specific intracellular inhibitor of Barnase, produced by the same organism [5]. Barstar can inhibit the intracellular enzymatic activity of Barnase through tight binding to this RNase, thus protect host cells from the damaging effect of Barnase [6].

It is well known that the formation of the Barnase-Barstar complex is mainly driven by electrostatic interactions [7]. Despite the well-recognized role of electrostatics in the formation and stability of Barnase-Barstar complex, the exact effects of changes in the electrostatics during the formation of the system, such as the presence of salt bridges, have not been examined thoroughly. To understand the influence of electrostatics in formation of Barnase-Barstar complex, we aim to analyze the electrostatic interaction behavior of the binding regions of this complex. We have simulated the Barnase-Barstar complex and Barnase alone under varying temperatures. The analysis specifically focuses on the contribution of salt bridges to the binding stability of the complex. Not only is this helpful for us to gain information about the roles of critical residues in the binding domain, but it also allows us to further understand the effects of electrostatic interactions on protein-protein complexes.

II. MATERIALS AND METHODS

The crystal structure of the Barnase-Barstar complex (1BRS) [8] that was obtained from the RSCB protein data bank (http://www.rcsb.org/) was used as a model for our experiment. A single complex (chain A and E) from the crystal structure was used for the complex simulation. As shown in Figure 1, chains A and E from the crystal structure were used for stand-alone protein simulations for Barnase and Barstar respectively.

The protein complex and stand-alone Barnase were subjected to molecular dynamics (MD) simulations using NAMD software [9]. The CHARMM27 force field [10] was used together with the MD simulations. Visual Molecular Dynamics (VMD) [11] was used to generate PSF files for the complex and proteins, in addition to visualization and data analysis. The cubic water boxes which contained TIP3P water molecules [12] were used to solvate the complex and protein. The simulated temperature was set to 300 K, 400 K, 500 K and 600 K, respectively. Through these temperatures, the protein can be observed from the natural to unfolded state. The short-range non-bonded interactions were calculated

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00063

using a 12 Å cut-off distance. The long-range electrostatic interactions were calculated by the Particle Mesh Ewald (PME) [13] method. All the bonds involving hydrogen atoms were constrained using the SHAKE method [14]. Firstly, the systems were performed with 10000 steps of steepest descent to finish energy minimization. Then, simulations were performed in NVT ensemble in which the number of particles, N, the volume, V, and the temperature, T, were kept constant. The time for each simulation was set to 10 ns and the time step was set to 2 fs.

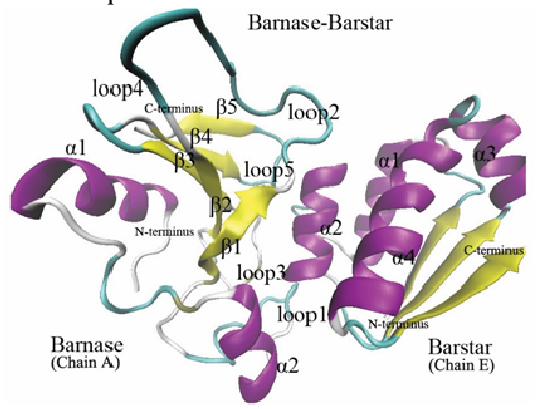


FIG.1. A cartoon representation of the structure of Barnase-Barstar complex

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. The Barnase-Barstar complex confers better structural stability at high temperatures

With molecular dynamics simulation, we probed the stabilities of the Barnase-Barstar complex and the individual protein Barnase. Here, we calculated the Cα root mean square deviation (RMSD) values of the structures in our simulations. The RMSD is a numerical measure to quantify differences between two conformations. It is defined as:

$$RMSD(t_j) = \sqrt{\frac{1}{N_i}\sum_{i=1}^{N_i}(r_i(t_j) - r_0)^2} \qquad (1)$$

where N_i is the number of atoms whose positions are being compared, $r_i(t_j)$ is the position of an atom i at time t_j, and r_0 is the reference value which takes the first time step of the simulation. In general, the RMSD value reflects the mobility of an atom during the molecular dynamics simulation trajectory. A higher residue RMSD value corresponds to higher mobility; conversely, a lower residue RMSD value indicates lower mobility.

The trend of the curves in Fig.2 shows that the systems were equilibrated and suitable to observe the dynamics of the model. In the 300K simulation, the complex and the separate Barnase were very stable throughout the simulation time. Upon inspection of RMSD plots, the average RMSD was found to be 1.64 Å for the Barnase in the complex, and the average RMSD was 1.84 Å for the independent Barnase in the 400 K simulation. In the trajectory run at 500 K, RMSD values of the Barnase in complex clearly increased from the initial conformation, which fluctuated between 0.56 Å and 3.81 Å; the average value of RMSD was 2.43 Å. For Barnase, RMSD values fluctuated between 0.58 Å and 4.57 Å, and the average value was 3.29 Å. When the temperature was set to 600 K, the plots of both the complex and Barnase indicated a clearly increasing trend, with the curves fluctuating more than other lower temperatures. In comparison to the Barnase in complex, with a RMSD value less than 7 Å, the RMSD values of independent Barnase increased significantly up to 12.29 Å. In summary, the RMSD diagrams show the stability of the molecular dynamics simulation trajectories, and they also indicate whether the protein can maintain a natural folding state at a given simulated temperature. When the protein tertiary structure begins to be disrupted, the RMSD value also rises to a higher value.

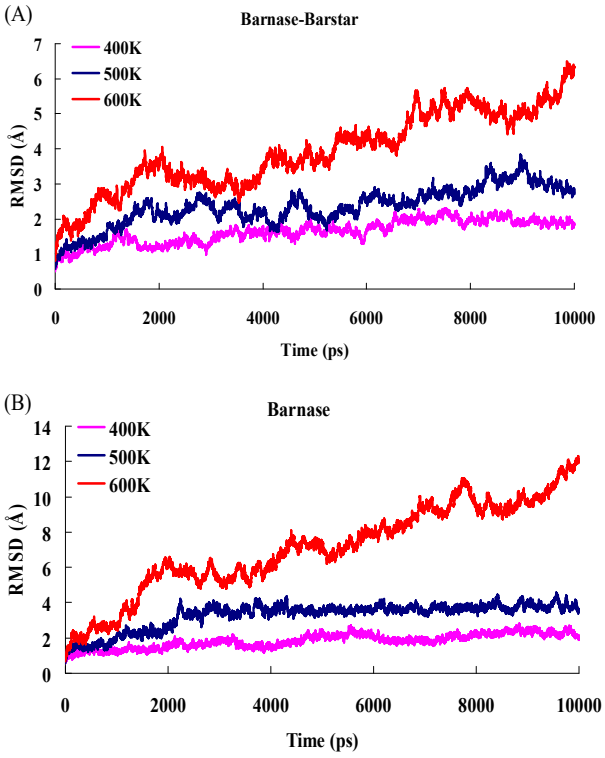


FIG.2 A-B. Root mean square deviation (RMSD) and time evolution plots of the Barnase-Barstar complex and Barnase at different temperatures ((A): Barnase-Barstar; (B): Barnase).

B. Influence of temperature on electrostatic interaction

In the Barnase-Barstar complex, electrostatic complementary regions are important factors that influence the association process and contribute to the stability of the

complex [15]. If charged residues in proteins are located within a certain distance, these residues can form salt bridges. As important contacts in electrostatic interactions, salt bridges have a crucial role in protein stability. To obtain information about the contributions of individual amino acids to complex association, we analyzed the course of salt bridge interactions in the binding regions of the complex. The last 5 ns simulation trajectory was used to analyze the fluctuations of the salt bridges.

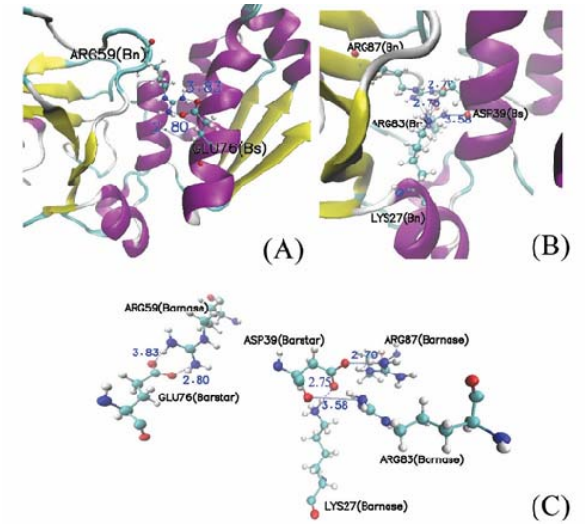


FIG.3 A-C. An overview of the location of salt bridges in the binding domain of the complex. A. The salt bridge between ARG59 and GLU76. B. The salt bridges between LYS27, ARG83, ARG87, and ASP39. C. A close demonstration of the four salt bridges independent of other residues. The blue dotted line represents a salt bridge interaction, and the number next to the dotted line represents the salt bridge distance. Red, oxygen; blue, nitrogen; cyan, carbon; white, hydrogen. (Bn: Barnase; Bs: Barstar)

In the Barnase-Barstar complex, four salt bridges, ARG59-GLU76, ARG83-ASP39, ARG87-ASP39 and LYS27-ASP39, are formed in the binding regions. Fig.3 shows the location of four salt bridges in the binding region. The occupancy of these salt bridges was analyzed in detail to estimate the change of electrostatic interaction. Fig.4 shows the change of salt bridge distance in the last 5 ns of the simulations. In a dynamic simulation at 300 K, these salt bridges were all found to be stable. When the simulation temperature was 400 K, ARG83-ASP39 and ARG87-ASP39 were found to be the most stable among these salt bridges; they were maintained within 4.0 Å distance without any separation. The average distances of ARG83-ASP39 and ARG87-ASP39 were 3.47 Å and 3.51 Å, respectively. In contrast, the salt bridge of ARG59-GLU76 experienced a short separation of approximately 1 ns, but it was stable at approximately 4.0 Å distance the rest of the time. Another salt bridge, LYS27-ASP39, was maintained within a short distance only in the first 1ns of the simulation. Later, the distance of the salt bridge was always more than 5.0 Å. The average distances of salt bridge LYS27-ASP39 were 3.78 Å and 5.56 Å, corresponding to the 300 K and 400 K simulations.

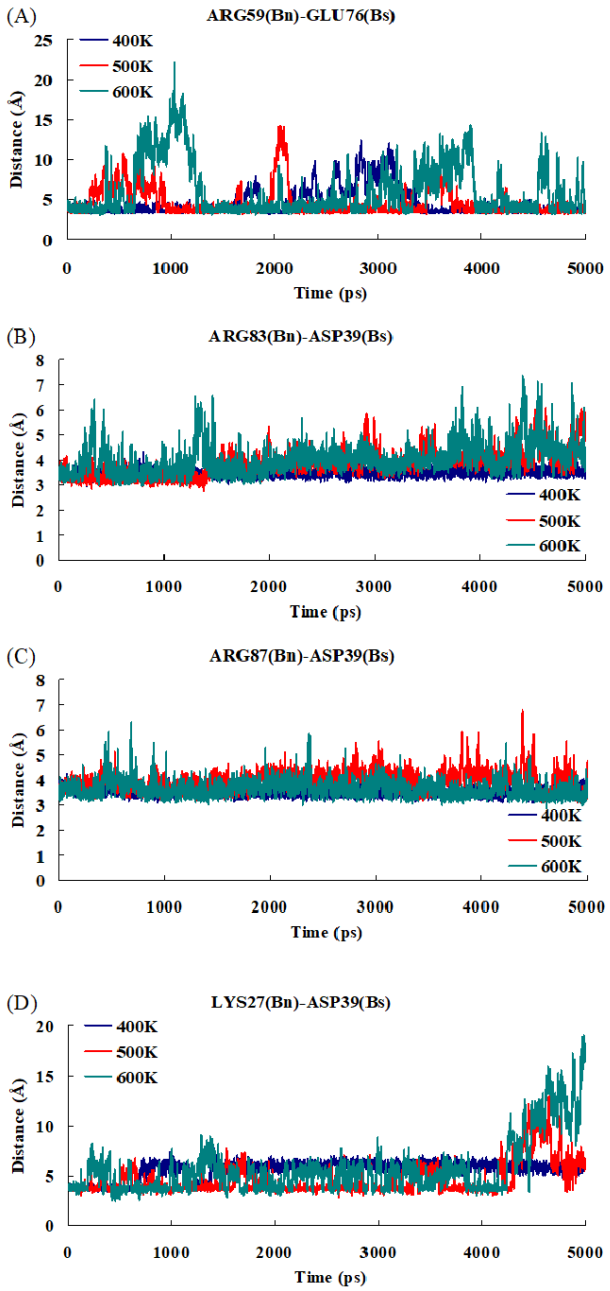


FIG.4 A-D. Plot of salt bridge distance changes as a function of time in the Barnase-Barstar complex binding domain at different temperatures.

Along with the rise of the simulation temperature, the salt bridge distances changed accordingly. In the 500 K

simulation, salt bridges ARG83-ASP39 and ARG87-ASP39 were still stable. Their average distances were 3.82 Å and 3.95 Å, respectively. ARG59-GLU76 was maintained within 4.0 Å for most of the simulation time, with only transient separation of the two residues. At 2061 ps, the distance of the salt bridge was up to 14 Å. However, when the simulation time was 2139 ps, the distance fell back rapidly to 4.08 Å. Salt bridge LYS27-ASP39 was also quite steady except for occasional separation. In general, the salt bridge LYS27-ASP39 in the 500 K simulation was more stable than in the 400 K simulation.

When the temperature increased to 600 K, among the above four salt bridges, ARG87-ASP39 demonstrated the most stable connection. This outcome may be related to the positions of the residues that form the salt bridge; the position of residue ASP39 is located in the α-helix of Barstar, while residue ARG87 is located in the β-sheet of Barnase. Both α-helix and β-sheet are relatively stable in the protein secondary structures. Meanwhile, there is another factor contributing to the stability of the salt bridge. In addition to the ARG87-ASP39 salt bridge, ASP39 can also form interactions with ARG83 and LYS27. In other words, the salt bridge network has a positive role in the relatively high stability of this salt bridge. For salt bridge ARG59-GLU76, ruptures and restorations were observed repeatedly during the whole simulation process.

On the basis of the above results, the distances of four salt bridges were less than 4 Å for most of the time from the 300 K to 500 K simulations. Namely, four salt bridges were well stabilized. The binding regions including these salt bridges were also stable. Thus, these salt bridges contribute to the stability of the complex in terms of binding interactions.

IV. Conclusion

Electrostatic interactions are a significant driving force for biological function. Clarifying the role of electrostatic interactions in protein function and stability is the key for protein engineering and drug design. In this article, we have analyzed electrostatic interactions of the Barnase-Barstar binding region, specifically the stability of salt bridge at different temperatures, using MD simulations. The side-by-side comparison between the complex and protein alone allows us to characterize the main binding features of the complex. Because of the electrostatic interactions in the binding domain, the unfolding temperature of Barnase in complex is higher than that of the independent enzyme. Barnase binds to Barstar residues that are located on a continuous polypeptide fragment; thus the binding interactions could prompt thermal stability of Barnase. Our results also clearly show that salt bridges are of major importance in protein stability at elevated temperature.

This article analyzed the contribution of residues in binding regions to the association of Barnase-Barstar. The results can be served as a reference for designing new thermostable Barnase mutant proteins according to electrostatic interactions. This type of analysis can also be used as an example to understand the role of electrostatic interactions for protein design.

Acknowledgment

This work is supported by the Qing Lan Project of Jiangsu Province (Year 2020 and Year 2019), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant Nos. 17KJB520039 and Nos. 18KJB520046, the Jiangsu Overseas Visiting Scholar Program for University Prominent Young and Mid-aged Teachers and Presidents.

References

[1] Francois M, Donovan P, Fontaine F. Modulating transcription factor activity: Interfering with protein-protein interaction networks. Seminars in Cell & Developmental Biology, 2020, 99: 12-19.

[2] Goodacre N, Devkota P, Bae E, Wuchty S, Uetz P. Protein-protein interactions of human viruses. Seminars in Cell & Developmental Biology, 2020, 99: 31-39.

[3] Portillo F, Vázquez J, Pajares MA. Protein-protein interactions involving enzymes of the mammalian methionine and homocysteine metabolism. Biochimie, 2020, 173: 33-47.

[4] Kostyukevich Y, Shulga AA, Kononikhin A, et al. CID fragmentation, H/D exchange and supermetallization of Barnase-Barstar complex. Scientific Reports, 2017, 7(1):6176.

[5] Alemany A, Rey-serra B, Frutos S, et al. Mechanical Folding and Unfolding of Protein Barnase at the Single-Molecule Level. Biophysical Journal, 2016, 110(1):63-74.

[6] Neumann J, Gottschalk KE. The Effect of Different Force Applications on the Protein-Protein Complex Barnase-Barstar. Biophysical Journal, 2009, 97(6):1687-1699.

[7] Kieslich CA, Jr RDG, Morikis D. Is the rigid-body assumption reasonable? Insights into the effects of dynamics on the electrostatic analysis of barnase–barstar. Journal of Non-Crystalline Solids, 2011, 357(2):707-716.

[8] Buckle A M, Schreiber G, Fersht A R. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0Å resolution. Biochemistry, 1994, 33(30):8878-8889.

[9] Phillips JC, Braun R, Wang W, et al. Scalable molecular dynamics with NAMD. Journal of Computational Chemistry, 2005, 26(16):1781-1802.

[10] MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins and nucleic acids. The Journal of Physical Chemistry, 1998, 102:3586-3616.

[11] Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. Journal of Molecular Graphics, 1996, 14(1):33-38.

[12] Jorgensen WL, Chandrasekhar J, Madura JD, et al. Comparison of simple potential functions for simulating liquid water. Journal of Chemical Physics, 1998, 79(2):926-935.

[13] Darden T, York D, Pedersen L. Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. Journal of Chemical Physics, 1998, 98(12):10089-10092.

[14] Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. Journal of Computational Physics, 1977, 23(3):327-341.

[15] Mitkevich VA, Schulga AA, Ermolyuk YS, et al. Thermodynamics of denaturation of complexes of barnase and binase with barstar. Biophysical Chemistry, 2003, 105(3):383-390.

Research on the Method of Leakage Inductance Suppression of High Frequency Transformer of Switching Power

LIU Jianbin
Jiangsu Key Lab of IoT Application Technology
Wuxi Taihu University
Wuxi, China
001210@wxu.edu.cn

Li Sha
Jiangsu Key Lab of IoT Application Technology
Wuxi Taihu University
Wuxi, China
lis@wxu.edu.cn

***Abstract*—Switching power supply is widely used in electrical systems. Its conversion efficiency is one of the key parameters for evaluating the quality of the power supply. At present, high frequency switching power supply is highly required with possible reduced size on transformers and inductors. When the working frequency reaches 10MHz in the switching power supply, the leakage inductance of high frequency transformer is the main factor affecting the overall conversion efficiency. By analyzing the main reason and the possible damage of the leakage inductance in high frequency transformer, a novel design is proposed in the paper, in which the primary and the secondary couplings are increased by the new design of the transformer, and the influence of the leakage inductance is suppressed by the RCD(Residual Current Device) clamp circuit. The optimized design is carried out on the 5V-2A low-power switching power which is widely used. The conversion efficiency of switching power is increased from 75.8% to 85.3%, and the withstand voltage of MOSFET(Metal-Oxide-Semiconductor Field-Effect Transistor) is reduced from 640V to 500V.**

Keywords-high frequency transformer; leakage inductance; electromagnetic coupling; clamp circuit

I. INTRODUCTION

Switching power is widely used. Its working principle is to control electronic switching devices (transistors, field-effect transistors, thyristors, etc.) to turn on and off continuously at a high frequency, and pulse modulate the input voltage, so as to realize AC(Alternating Current) / DC(Direct Current), DC / AC, DC / DC and other voltage conversion, as well as the adjustable control of output voltage and current. The core technical index of switching power is conversion efficiency η, which refers to the proportion of output power to input power, and the difference between input power and output power is transformed into heat energy emission. Obviously, improving the conversion efficiency can improve energy utilization efficiency, energy conservation and emission reduction, and reduce energy consumption; on the other hand, it can reduce the heat generated by power loss, improve the working environment of switching power and delay its life.

To improve the switching power conversion efficiency η, the core method is to increase the working frequency. At present, power electronic technology has made the working frequency exceed 10MHz, so that the conversion efficiency can reach more than 95% [1].With the increase of switching power frequency, the influence of high frequency transformer leakage inductance on conversion efficiency becomes more and more prominent, which becomes the main factor to restrain the further improvement of conversion efficiency. Optimizing the design and suppressing the leakage inductance of high-frequency transformer has become an effective path to improve the switching power conversion efficiency η and reduce the failure rate [2].

II. LEAKAGE INDUCTANCE AND ITS INFLUENCE

High frequency transformer is a device for energy storage, release and conversion. It converts the input electric energy into magnetic energy and stores it in the magnetic core, and then releases the magnetic energy to the output end in the form of electric energy. In the process of transformer manufacturing, the primary and secondary coils can not achieve the ideal coupling degree, and the magnetic circuit of the magnetic core also needs to reserve a variable air gap to adjust the inductance. Therefore, the magnetic line generated by the primary coil can not be coupled to the secondary coil through the magnetic core, and a part of the magnetic flux will be leaked into the air to form a closed magnetic circuit. The leakage magnetic flux is called magnetic-flux-leakage, which produces leakage inductance.

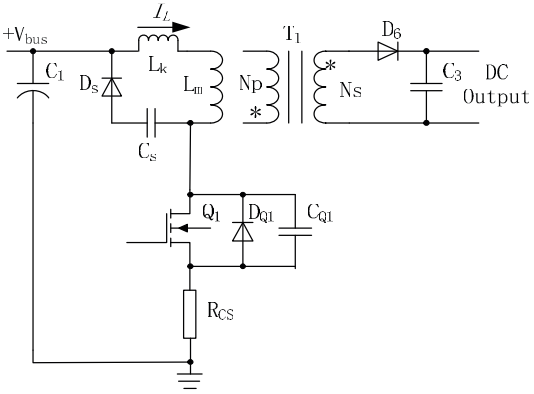


Figure 1. Diagram of main circuit of transformer

Figure 1 shows the diagram of high-frequency transformer[3-5].Transformer T_1 includes ideal transformer composed of primary winding N_p and secondary winding N_s, primary side excitation inductance L_m, leakage inductance L_k, parasitic diode D_{Q1} of Q_1, parasitic capacitance C_{Q1}, absorption capacitor C_s, absorption diode D_s, input rectified bus voltage V_{bus}, output voltage V_o.

The calculation formula of energy storage of high frequency transformer core is as follows:

$$J = \frac{1}{2} \times L_P \times I_{PK}^2 \quad (1)$$

The excitation inductance L_m and leakage inductance L_k are connected in series, and the flowing current are the same:

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00064

$$i_{Lm} = i_{Lk} = \frac{V_{bus}}{L_k + L_m} \times t \quad (2)$$

The stored energy of inductance L_m and leakage inductance L_k are obtained by substituting equation 1 and equation 2:

$$J_{Lm} = \frac{1}{2} \times L_m \times I_{Lm}^2 \quad (3)$$

$$J_{Lk} = \frac{1}{2} \times L_k \times I_{Lk}^2 \quad (4)$$

The energy storage of excitation inductor L_m is coupled with the secondary and can be transmitted to the secondary output. The energy storage of leakage inductance L_k can not be coupled to the secondary, and it will be consumed in the oscillation with the parasitic capacitance of the line, which will increase the loss and decrease the conversion efficiency. Excitation inductance L_m and leakage inductance L_k oscillate with parasitic capacitance C_{Q1} respectively. The V_{ds} waveform generated after MOSFET switch-off is shown in Figure 2:

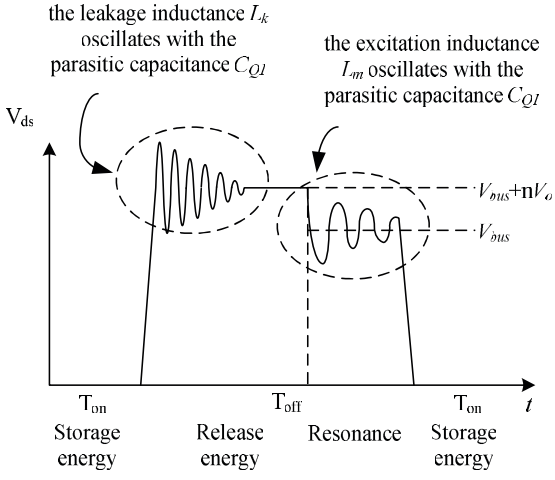


Figure 2. V_{ds} waveform after MOSFET switch-off

During the MOSFET turn-on period (T_{on}), the transformer energy storage, V_{ds} is 0V. When MOSFET is turned off (T_{off}), the transformer releases the stored energy to the secondary. At the same time, the leakage inductance L_k oscillates with the parasitic capacitance C_{Q1}. The oscillation center voltage is V_{bus}+nV_o (n = N_p / N_s). The amplitude of leakage inductance oscillation is proportional to leakage inductance L_k and parasitic capacitance C_{Q1}. It can be seen that the larger the leakage inductance L_k, the greater the oscillation energy loss, the lower the conversion efficiency, and the greater the overvoltage damage to MOSFET. When the primary energy storage is released, the excitation inductance L_m and parasitic capacitance C_{Q1} oscillate. The upper limit of the oscillation is V_{bus}+nV_o, and the energy loss and damage to MOSFET are lower than that of leakage inductance L_k.

The existence of leakage inductance L_k will affect the performance of switching power. Especially, with the increase of the converter frequency f_{sw}, the permeability of ferrite core material of transformer will decrease, and the excitation inductance L_m will decrease. However, the leakage inductance L_k is affected by the transformer process, and the leakage inductance value remains unchanged, so the relative value of the excitation inductance L_m increases, and the influence on the power converter becomes greater. At present, the working frequency of AC / DC converter can reach 10MHz, and leakage inductance is the main factor affecting conversion efficiency.

III. METHOD OF LEAKAGE INDUCTANCE SUPPRESSION

To suppress the leakage inductance and improve the switching power conversion efficiency η is mainly implemented in two aspects: optimizing the transformer design to reduce the leakage inductance [6], increasing clamp circuit to suppress the influence of leakage inductance [7].

A. Optimization of transformer design

The transformer in this design is shown in T_1 in Figure 1. The parameters of the transformer are calculated as follows: primary winding N_p is 96 turns, secondary winding N_s is 5 turns, auxiliary winding N_a is 11 turns, excitation inductance L_m is 2mH, EE13 vertical 5+5 pins skeleton and PC40 magnetic core are selected.

The conventional winding method is shown in Figure 3, with the core and framework as the core, winding outwards in turn: primary winding N_p (96 turns, wire diameter 0.2mm), secondary winding N_s (5 turns, twisted pair, wire diameter 0.45mm), auxiliary winding N_a (11 turns, wire diameter 0.2mm). The star is homonymous end, and the arrow is the in and out direction of enameled wire.

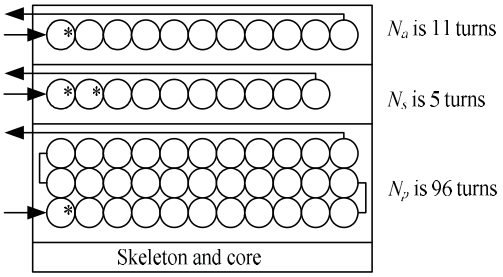


Figure 3. Conventional winding method of transformer

After testing, the transformer primary leakage inductance L_k is 220uH, accounting for 11% of the excitation inductance L_m is 2mH, and the conversion efficiency η is 75.8%. In order to suppress the leakage inductance and improve the conversion efficiency, the transformer is improved as follows:

- Sandwich winding method is adopted, that is, each winding is interleaved to strengthen coupling.
- Each group of windings should be wound tightly and evenly distributed with no gap.
- The outgoing line should be at right angle as far as possible and close to the skeleton wall.
- If the last layer can't be wrapped with one layer, it will be evenly wrapped.
- The thickness of the insulation layer should be as thin as possible to meet the withstand voltage without leaving gaps.

The improved winding method is shown in Figure4. After testing, the transformer primary leakage inductance L_k is 16uH, accounting for 0.8% of the excitation inductance L_m is 2mH, and the conversion efficiency η is 79.2%, which is 3.4% higher than the conventional winding method.

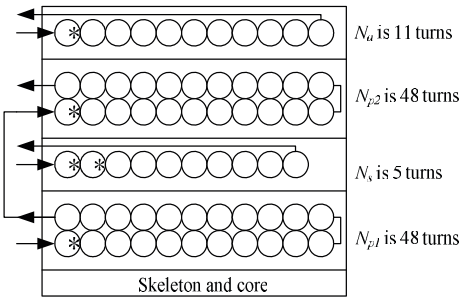


Figure 4. Low leakage inductance winding method of transformer

B. *Add RCD clamp circuit*

The leakage inductance L_k oscillates with the parasitic capacitance C_{Q1}, and the energy can not be coupled to the secondary winding N_s, which needs to be consumed gradually in the oscillation. Obviously, the larger the amplitude of oscillation is, the more energy is consumed, and the lower the conversion efficiency η. Clamp circuit can be used to suppress the oscillation caused by leakage inductance.

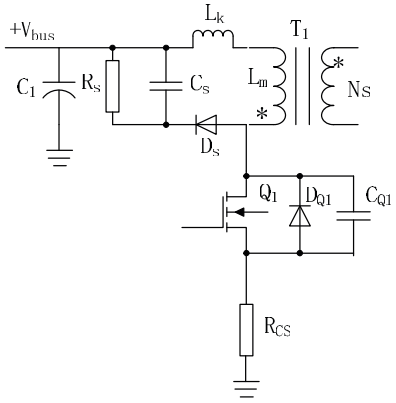


Figure 5. RCD clamp circuit

As shown in Fig.5, R_s, C_s and D_s constitute the RCD clamp circuit, which is used to absorb the oscillation generated by leakage inductance L_k and parasitic capacitance C_{Q1}. Diode D_s controls the working current direction of RCD. When Q_1 is on, D_s is cut off and RCD is not working. When Q_1 is off, leakage inductance L_k oscillates with parasitic capacitance C_{Q1}, and D_s is on to provide free wheeling channel to charge C_s, so that C_s can absorb the oscillation generated and limit the amplitude of oscillation voltage. When Q_1 is on and RCD is not working, the absorption resistance R_s releases the energy absorbed by C_s, which is ready for the next RCD cycle.

IV. VERIFICATION AND TESTING

This design chooses to implement improvement measures on the scheme of low power switching power with input of 220V and output of 5V-2A. This kind of switching power is commonly used in small household appliances, such as adapter, charger and so on. It has the most extensive use and the largest demand. Therefore, it is of great significance to improve its conversion efficiency. Due to the small power, the conversion efficiency of this kind of switching power is relatively low, and the conversion efficiency η is about 80%. The sample of switching power is shown in Figure 6 below:

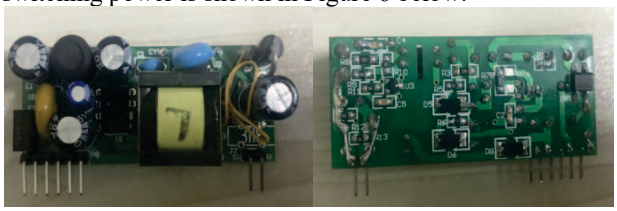


Figure 6. Photo of sample static display

The improvement is carried out on the sample, and the comparison test is carried out to verify the implementation effect of the improved transformer process to reduce the leakage inductance L_k and increase RCD clamping circuit to suppress the leakage inductance hazard.

A. *High leakage inductance and non clamping circuit*

The transformer scheme without improvement is shown in Fig.3, the leakage inductance of primary side L_k is 220uH, without RCD clamp circuit. Input 220V, output 5V-2V, D-S(Drain-Source) waveform V_{ds} of Q_1 in oscilloscope is shown in Figure 7 below. The results show that the peak value of V_{ds} is 640V, the oscillation amplitude is 360V, and the conversion efficiency of converter is 75.8%.

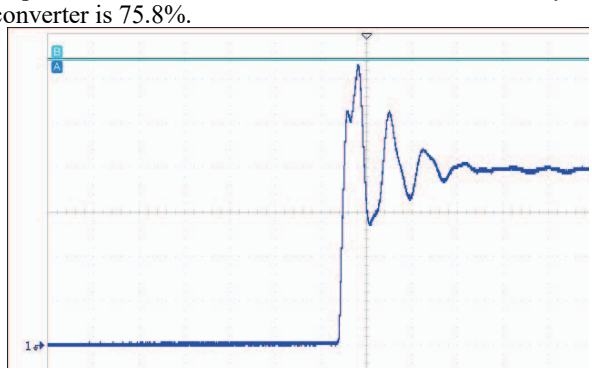


Figure 7. V_{ds} waveform of high L_k and non clamping circuit

B. *Low leakage inductance and non clamping circuit*

The improved transformer scheme as shown in Fig. 4 is adopted, the leakage inductance of primary side L_k is 16uH, without RCD clamp circuit. Input 220V, output 5V-2A, D-S waveform V_{ds} of Q_1 in oscilloscope is shown in Figure 8. The peak value of V_{ds} is 560V, and the oscillation amplitude is 240V. After calculation, the conversion efficiency η of converter is 79.2%.

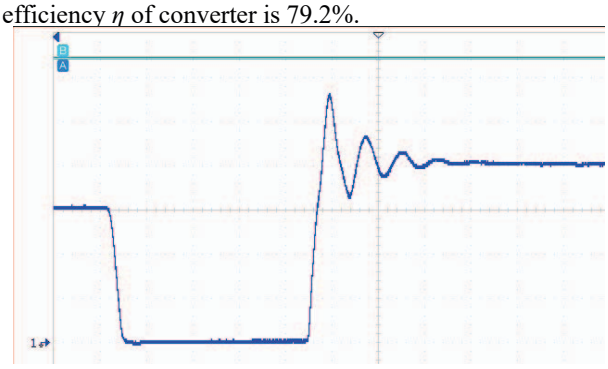


Figure 8. V_{ds} waveform of low L_k and non clamping circuit

C. *High leakage inductance and clamping circuit*

The scheme of no improved transformer as shown in Fig. 3 is adopted, the leakage inductance L_k of primary side is 220uH, and RCD clamp circuit is added. Input 220V, output 5V-2A, D-S waveform V_{ds} of Q_1 in oscilloscope is shown in Figure 9. The results show that the peak value of V_{ds} is 540V, the oscillation amplitude is 220 V, and the conversion efficiency of converter is 80.2%.

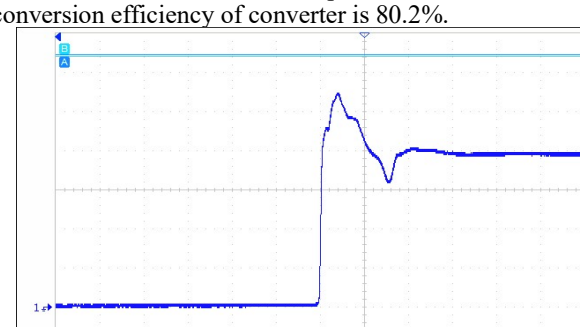


Figure 9. V_{ds} waveform of high L_k and clamping circuit

D. *Low leakage inductance and clamping circuit*

The improved transformer scheme as shown in Fig. 4 is adopted, the leakage inductance L_k is 16uH and RCD clamp circuit is added. Input 220V, output 5V-2A, D-S waveform V_{ds} of Q_1 in oscilloscope is shown in Figure 10. The results show that the peak value of V_{ds} is 500V, the oscillation amplitude is 120V, and the conversion efficiency of converter is 85.3%.

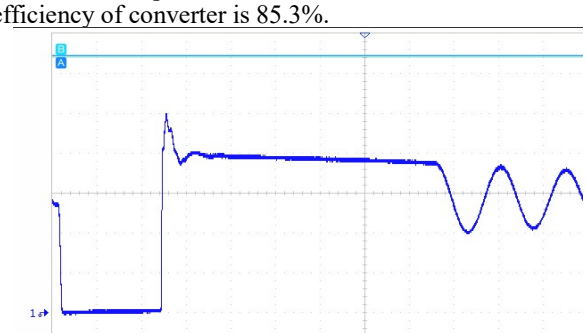


Figure 10. V_{ds} waveform of low L_k and clamping circuit

Table 1 is obtained by summarizing the data of Figure 7 to figure 10. The influence of leakage inductance and RCD clamp circuit on AC / DC converter is given.

TABLE I. COMPARISON OF EXPERIMENTAL RESULTS

leakage inductance L_k (uH)	clamp circuit	V_{ds} peak value (V)	oscillation amplitude (V)	conversion efficiency η
220	No	640	360	75.8%
16	No	560	240	79.2%
220	Yes	540	220	80.2%
16	Yes	500	120	85.3%

It can be seen from table 1 that by improving the transformer process to reduce the leakage inductance L_k and increasing the RCD clamp circuit can effectively reduce the withstand voltage of MOSFET V_{ds} and improve the converter conversion efficiency η. The RCD clamp circuit can also reduce the oscillation times and periods. In comparison, the influence of RCD clamp circuit is slightly greater than that of leakage inductance. In practical application, it is necessary to reduce leakage inductance L_k and increase RCD clamp circuit at the same time.

V. CONCLUSIONS

This paper analyzes the cause and influence of leakage inductance of high frequency transformer. Under the condition that the working frequency of AC / DC converter has reached 10MHz, leakage inductance has become the main factor affecting conversion efficiency. In this paper, two improvement methods are adopted. One is to optimize the transformer design, such as sandwich winding method, reducing winding gap, winding evenly distributed winding, reducing the thickness of insulation layer and so on, so as to increase the coupling of primary and secondary transformer and reduce the generation of leakage inductance. The second is to add RCD clamping circuit to suppress the influence of leakage inductance. The conversion efficiency is gradually increased from 75.8%, 79.2%, 80.2% and 85.3%, and the V_{ds} withstand voltage of MOSFET is gradually reduced from 640V, 560V, 540V and 500V, which shows that the improved method is effective.

ACKNOWLEDGMENT

The authors thank all partners in Jiangsu Key Lab of IoT Application Technology (Wuxi Taihu University). This research was supported in part by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 18KJB510045), and in part by the Research Foundation of Jiangsu Key Construction Laboratory of IoT Application Technology (Grant No. 18WXWL05).

REFERENCES

[1] Song Qiang, Zhao Biao, Liu Wenhua, et al. Next generation high-frequency-isolation power conversion technology for smart grid[J]. Proceedings of the CSEE, 2014, 34(36): 6369-6379.(in Chineses)
宋强, 赵彪, 刘文华, 等. 智能电网中的新一代高频隔离功率转换技术[J]. 中国电机工程学报, 2014, 34(36): 6369-6379.

[2] Ren Lei, Wei Zheng, Gong Chunying, et al. Fault feature extraction techniques for power devices in power electronic converters a review[J]. Proceeding of the CSEE, 2015, 35(12): 3089-3101.(in Chineses)
任磊, 韦徵, 龚春英, 等. 电力电子电路功率器件故障特征参数提取技术综述[J]. 中国电机工程学报, 2015, 35(12): 3089-3101.

[3] Abed N Y, Mohammed O A. Physics-based high-frequency transformer modeling by finite elements[J]. IEEE Transaction. on Magn, 2010, 46(8): 3249-3252.

[4] Jazebi S, Zirka S E, Lambert M, et al. Duality derived transformer models for low-frequency electromagnetic transients part I: topological models[J]. IEEE Transactions on Power Delivery, 2016, 31(5): 2410-2419.

[5] Dujic D, Mester A, Chaudhuri T, et al. Laboratory scale prototype of a power electronic transformer for traction applications[C]. Proceeding of the 2011-14th European Conference on Power Electronic and Applications, 2011: 1-10.

[6] Deng Xianming, Cheng En, Liu Na. A novel transformer with adjustable leakage inductance[J]. International Journal of Applied Electromagnetics and Mechanics, 2017, 55(S1):1-14.

[7] Huang Yiyi, Ben Hongqi, Wang Weiqi, Zhou Di, Qi Chao. Active Clamp Circuit Applying to Absorbing Voltage Spikes in Single-Stage Full-Bridge PFC Converter[J]. Transactions of China Electrotechnical Society,2018,33(04):845-855.(in Chineses)
黄奕豰, 贲洪奇, 王韦琦, 周迪, 齐超. 用于单级桥式 PFC 变换器电压尖峰抑制的有源钳位电路[J]. 电工技术学报, 2018,33(04):845-855.

2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Study on Carbon Sink Value of Marine Resources

Ying Zou
International Business Faculty, BNUZ
Zhuhai, China
Ying.zou@smu.ca

Xinyi Liu
Department of Management, BNUZ
Zhuhai, China
1473904165@qq.com

Liying Luo
Department of Pharmacy, CPU
Nanjing, China
157061083@qq.com

***Abstract*—Ecosystem service has two characteristics: one is typical public goods, the other is external, so the development and utilization of ecological environment cannot be realized directly through the market. It is necessary to take adequate measures to make up for the shortage of market mechanisms, so the evaluation of ecosystem service value has gradually become the focus of scholars. This paper uses conditions with Guangdong area as an example, evaluating the ocean carbon sink, the economic value of the services, the influencing factors of Guangdong residents' willingness to pay, it is concluded that Guangdong ocean carbon sinks the higher total monetary value of ecosystem services, the respondent's age, level of education, occupation, family income, degree of understanding of the ocean carbon sink, all the five factors have a significant influence on willingness to pay. Gender did not correlate with a willingness to pay.**

Keywords-Marine resources; Carbon sink value; Guangdong

I. INTRODUCTION

The ocean is the planet's most significant carbon sink, capturing carbon dioxide from burning fossil fuels and thus slowing the pace of climate change. Compared with forest carbon sinks, ocean carbon sinks have the advantages of more extensive carbon storage and longer storage time. The Marine carbon sink converts reactive organic carbon into inert organic carbon by the tiny but abundant micro-organisms in the ocean so that carbon can be stored in the sea for a long time.

Guangdong is rich in water resources, the total annual precipitation of 319.4 billion cubic meters, runoff volume of 181.9 billion cubic meters, plus the neighboring from the Xijiang and Hanjiang River, such as water volume of 233 billion cubic meters, in addition to the deep groundwater of 6 billion cubic meters, per capita water resources available for exploitation of 4735 cubic meters, significantly higher than the average level in China. Guangdong province has numerous rivers, mainly the Pearl River Basin (Dongjiang river, Xijiang River, Beijiang River, and Pearl River Delta Region), the Hanjiang River basin, and the coastal rivers in eastern and western Guangdong, with the catchment area accounting for 99.8% of the province's total area, and the other rivers in the Yangtze River basin are Poyang Lake and Dongting Lake. There are 542 tributaries at all levels with a basin area of more than 100 square kilometers.

II. SAMPLE DATA CHARACTERISTICS AND DATA DESCRIPTION

A. Socio-economic characteristics

Among the respondents, males accounted for a more significant proportion, accounting for 53.5% of the total, while females accounted for 46.5%.

The age of the respondents was mainly between 18 and 30 years old, accounting for 42.4 percent of the total number of respondents. The proportion of 31-40 years old is 30.3%. Those under 18 and those between 41 and 50 accounted for 9.8% and 9.4% of the total. Those aged 51-60 and those over 60 accounts for 5.8% and 2.3% of the total, respectively. The overall age structure of the respondents shows an inverted U-shaped distribution.

The educational level of the respondents covered all levels from primary school to graduate school, with high school education and college education accounting for 30.5% and 27.7% of the total number of respondents respectively, 23.5 percent of those with junior secondary education; Those with primary education or below accounted for 11.7%; Those with graduate degrees or above accounted for 6.6%, which was the most minor proportion.

The respondents' occupations covered a wide range, mainly students, farmers, and enterprise employees, accounting for 26.6%, 25.4% , and 19.2% of the total number of respondents, respectively.

The monthly family income of the investigated is mainly concentrated in the income level of 3000-5000 yuan, accounting for the highest proportion of 50.5%; The second is the income level of 1000-3000 yuan accounted for 26.2% of the overall income level of the survey; The proportion of 5,000-10,000 yuan was 10.9%; The ratio of the public below $1,000 was 9.8%; Those with more than 10,000 yuan accounted for 2.6 percent of the total. It can be seen that in the overall monthly household income structure of the respondents, there are fewer low-income and high-income people, and more middle-income people.

III. EMPIRICAL ANALYSIS OF MARINE CARBON SINK VALUE ASSESSMENT IN THE STUDY AREA

B. Calculation of average willingness to pay

1）Nonparametric estimation of average willingness to pay:

Payment card guidance techniques can be used to reflect the range of willingness to pay variation,

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00065

depending on the maximum value that respondents are willing to pay and the value that they are explicitly unwilling to pay. In the analysis of samples, it is concluded that the maximum WTP scatter points of most respondents are distributed on the numbers 100, 125, 150, 200, 250, 300, and 350.

Based on the calculation formula of willingness to pay for the restoration of Ejin Banner ecosystem services by Xu Zhongmin, Zhang Zhiqiang, and Cheng Guodong, the public's expectation of positive willingness to pay for Marine carbon sink ecosystem services can be calculated as follows according to the frequency distribution of willingness to pay:

$$E(WTP) = \sum PiBi = 162.8(yuan) \quad (1)$$

In the formula, WTP represents the willingness to pay that the respondent accepts;

Pi indicates the frequency of the bid amount that respondents are willing to pay;

Bi said respondents would undoubtedly be willing to pay the bid amount.

Because the payment card guide technology, respondents are willing to pay the biggest value represents the lower limit of the overall average WTP, the respondents are opposed to paying than their chosen number higher level of value, the mathematical expectation formula, can get the limit, the overall average WTP is not willing to pay the minimum amount of:

$$E(NWTP) = \sum PiBi = 224.3(yuan) \quad (2)$$

Where, NWTP represents the value of willingness to pay rejected by respondents.

The maximum acceptable amount of public willingness to pay was 162.8 yuan, and the number of unwillingness to pay was 224.3 yuan. Therefore, it can be estimated that the average desire of Guangdong residents to pay for Marine carbon sequestration services is 162.8 ~ 224.3 yuan.

2）EstimatioParameter estimation method of average willingness to pay

Using payment card guidance technology can directly reflect the WTP range of respondents. Once the variation range of the average WTP is clear, the parameter estimation method can determine the specific value of the average willingness to pay. The parameter estimation method is adopted, which must establish a functional relationship with the interviewees' willingness to pay and the attribute value of public goods. The median bid range is taken as the explanatory variable. The maximum likelihood estimation method is used to estimate the basic information and attribute variables of the median bidding quantity and other ecological services. In the specific application, the lognormal distribution of WTP is generally selected as the explanatory variable.The equation is as follows:

$$\ln(WTP) = X'\beta + u \quad (3)$$

X represents some characteristic attributes of the respondent or ecological, environmental service, and X' represents its transpose matrix; β represents the estimated parameter;

U is the random variable that follows a normal distribution$[0, \sigma^2]$.

Many scholars found in later studies that the assumption in Equation (3) that the expected WTP value is equal to the midpoint of the variation range of the number of bids made by respondents may lead to biased WTP value. Cameron and Huppert treated WTP as a random variable based on Equation (3) and introduced a double-boundary probability model for analysis.

After investigating the range of WTP using payment card guidance technology, respondents expressed the probability of answering 'yes' as follows:

$$Pr('Yes') = Pr(WTP \geq T_L) = 1 - GWTP(T_L) \quad (4)$$

The probability that WTP falls between$[T_L, T_U]$ the bidding thresholds is $GWTP(T_U) - GWTP(TSL)$, where T_L is the maximum value interviewees are willing to pay, and T_U is the minimum value interviewees are not willing to pay. Obviously, the respondents' willingness to pay falls within this range$[T_L, T_U]$. For all participants, the following logarithmic likelihood function can be obtained:

$$\ln(L) = \sum \ln[GWTP(T_{Ui}) - GWTP(T_{Li})] \quad (5)$$

C. Elimination of the total economic value of carbon sinks in Guangdong, Hong Kong, and Macao

According to the above calculation, the average annual amount that sample residents in Guangdong are willing to pay for Marine carbon sequestration services is 183.2 yuan per household. In the 600 questionnaires, the positive rate of willingness to pay is 83.67%, and it can be concluded that the average annual willingness to pay of Guangdong residents is 153.28 yuan/household. Guangdong Province has 42.4692 million households and 4.2225 million collective families, with a population of 111671837 and 14340673, respectively. The average population per household was 2.63 people, 0.57 less than the 3.20 people in the sixth national census in 2010. The total number of households is 46.6916 million, so the number of permanent families willing to pay is 39.9669 million.

Combined with existing studies, this paper intends to measure the value of Guangdong's Marine carbon aggregation by using the annual average willingness to pay (WTP) of Guangdong residents. The calculation formula is as follows:

Total WTP of Residents in Guangdong = average annual WTP of residents × number of permanent households WTP in Guangdong（6）

Substitute in Equation (6), WTP of Guangdong residents to ocean carbon sink is 5.99×109 yuan/year. According to the statistical results of sample payment, the average monthly income of the respondent's family is about 13,512 yuan, and the average annual income is 162,144 yuan. According to this value, the respondents' average willingness to pay is about 0.26 percent of yearly household income. This shows that the Marine carbon

sequestration fund project is within the reasonable range that residents can bear and will not bring a heavy economic burden to families.

D. LOGISTIC regression analysis of willingness to pay

1) Selection and interpretation of variables

All samples in this study are derived from the field research data to evaluate of the Marine carbon sink value in Guangdong in 2021. In this paper, based on the previous studies on conditional value assessment and case analysis, combined with the actual situation of Guangdong, variables affecting the public's willingness to pay for Marine carbon sequestration services are mainly classified into three categories. One is the individual factors of respondents, such as gender, age, and education level; Second, family status, primarily refers to family income; The third is the cognition of Marine carbon sink ecological services. This study selected six possible evaluation factors that may have an impact on the willingness to pay for Marine carbon sink services, and the description of each variable is shown in Table 1:

TABLE I. SELECTION AND EXPLANATION OF VARIABLES

types of variables	variable name	Variable to explain
decision variable	Whether respondents are willing to pay a certain amount for Marine carbon sink ecological services (Y)	Yes =1, no =2
personal factors	Gender (x1*)	Male =1, female =2
	Age (x2*)	——
	Education level (x3*)	Primary school and below =1 Junior high school = 2 High school = 3 University = 4 Graduate students and above =5
	Occupation (X4 *)	Students = 1 Workers = 2 Farmers = 3 Administrative personnel =4 Employee =5 Unemployment = 6 Housewife =7 Self-employed =8 Other = 9
family status	Monthly household income (X5 *)	——
Knowledge of Marine carbon sequestration services	Understanding of carbon sinks (x6*)	I don't know anything about =1 Heard of = 2 We know that theta is equal to 3 Fully understand =4

Note: the (*) in the variable name represents a dummy variable

2) Construction of Logistic regression model

Logistic regression is a regression analysis of qualitative variables, and the Logistic model is one of the commonly used models to study residents' willingness to pay. Logistic regression is an extension of multiple regression analysis, and its dependent variable is a discontinuous variable. This study takes Guangdong residents' willingness to pay for Marine carbon sequestration services as the explained variable. It conducts a Logistic regression analysis on the factors that may affect residents' consent or disapproval to pay. Assign the answer "willing to pay" to 1 and "unwilling to pay" to 0. These factors that may affect the willingness to pay come from the socio-economic characteristics and personal information of the survey sample, such as age, sex, education level, monthly family income, occupation, and the respondents' familiarity with the ocean carbon sink. The Logistic regression model is as follows:

$$\mathrm{Log}[\mathrm{Prob(Yes)}/\mathrm{Prob}(1-\mathrm{Yes})] = [\mathrm{B}_0 + \mathrm{B}_1(\mathrm{age}) + \mathrm{B}_2(\mathrm{sex}) + \mathrm{B}_3(\mathrm{education}) + \mathrm{B}_4(\mathrm{income}) + \mathrm{B}_5(\mathrm{profession}) + \mathrm{B}_6(\mathrm{familiar})] \quad (7)$$

Prob(Yes) represents the probability that residents are willing to pay, and its value ranges from 0 to 1.
Prob(1-yes) represents the probability that residents are not willing to pay;
B is the constant term;
B_1、B_2、B_3、B_4、B_5、B_6represents the regression coefficient of each variable;
Age indicates the age of the respondent;
Education represents the education level of the respondents;
Income refers to the monthly household income of the respondent;
Familiar indicated respondents' knowledge of Marine carbon sequestration services.

3) Model regression results and analysis

TABLE II. MODEL REGRESSION RESULTS

The independent variables	regression coefficient	significance level	χ^2
gender	0.021	0.023	8.463
age	0.016	0.023	10.371
Education Level	0.023	0.043	16.256
professional	0.026	0.037	17.328
Monthly household income	0.041	0.024	12.152
How well we know about ocean carbon sinks	0.163	0.008	10.214

From the above regression results, it can be seen that:
The value of B_1、B_2、B_3、B_4、B_5、B_6 directly reflects the sensitivity of the willingness to pay of the investigated samples.
Gender does not correlation with a willingness to pay.
The age factor in the model shows a significant effect on willingness to pay.

There is a significant favorable influence between the education level and the willingness to pay; the higher the education level, the greater the corresponding willingness to pay. This may be because more educated residents, who are more environmentally conscious and have higher incomes, can afford a higher willingness to pay.

The occupation of residents has a significant impact on their willingness to pay. The more stable their work is, the more concerned they are about environmental protection, and the more likely they are to have a higher willingness to pay.

There is a significant favorable influence between residents' income and willingness to pay. The higher the monthly income of a family, the greater its willingness to pay and the more able it is to pay.

There is a significant influence between residents' knowledge of ocean carbon sinks and their willingness to pay. With more understanding of ocean carbon sinks, residents will realize that ocean carbon sinks play an essential role in mitigating carbon content in the atmosphere and improving residents' living environment. Therefore, the residents are willing to pay a certain amount of money for this, and of course, the willingness to pay will be more vital.

Ⅳ. Conclusion

The main factors influencing WTP are age, household income, education level, and awareness of ocean carbon sink. Gender did not correlate with a willingness to pay. Income is positively correlated with WTP, and the desire to pay of high-income people is generally higher than that of low-income people. The higher the level of education, the greater the willingness to pay. The higher the understanding of ocean carbon sink, the higher the WTP of residents; On the other hand, residents who know little about WTP have fewer WTP and even zero payment. Under the urgent situation of global warming, accelerating emission reduction and increasing carbon sink is a significant issue of international concern, and ocean carbon sink is a critical way to solve the problem of increasing carbon sink effectively. In 2015, the CPC Central The overall plan for reforming the ecological civilization system promulgated by the State Council put forward the "establishment of an effective mechanism to increase marine carbon sequestration" to bring marine carbon sequestration into the national strategic level. In 2016, the country's "more tough choice-- more severe consequences-- in planning officials, marine carbon sequestration will be listed as one of the critical research and development, and oceanic carbon sequestration will be incorporated into the policy system China's economic and social development. It is imperative to accelerate the study of ocean carbon sink in China. The calculation of economic value of ocean carbon sink is the basis for promoting the development of ocean carbon sink projects and carrying out ocean carbon sink trading. It is necessary to continue to strengthen the research and exploration of relevant theories and practices, to provide technical support for China to lead the development of the international ocean carbon sink and respond to global climate change.

References

[1] Of environment D. Summary, including Headlines[J]. Department of the Environment.

[2] Saokhata J, Lakanavichianb S. Forest Carbon Sequestration in Huai Kong Ngong Sub-watershed, Samoeng District, Chiang Mai Province, Thailand.

[3] Atkin B, Gong T, Harmon J, et al. Carbon sequestration method:, US20110027017[P]. 2011.

[4] TimothyBeatley, et al. The Urban-Ocean Connection[M]. Island Press/Center for Resource Economics, 2014.

[5] Chen Y, Chen H Y, Jan S, et al. Phytoplankton productivity enhancement and assemblage change in the upstream Kuroshio after typhoons[J]. Marine Ecology Progress, 2009, 385:111-126.

[6] Burrows M, Hughes D, Austin W, et al. Assessment of Blue Carbon Resources in Scotland's Inshore Marine Protected Area Network. 2017.

[7] Mitra S, Bianchi T S. A preliminary assessment of polycyclic aromatic hydrocarbon distributions in the lower Mississippi River and Gulf of Mexico[J]. Marine Chemistry, 2003, 82(3–4):273-288.

[8] Zhang B. Carbon sink by marine fishing industry[J]. Progress in Fishery Sciences, 2013.

[9] Sukarni, Sudjito, Hamidi N, et al. Potential and properties of marine microalgae Nannochloropsis oculata as biomass fuel feedstock[J]. International Journal of Energy and Environmental Engineering, 2014, 5(4):279-290.

[10] Chun-Hou L I, Jia X P, Zhan-Hui Q I, et al. Effect Evaluation of a Low-carbon Fisheries Production by Marine Ranching in Daya Bay[J]. Journal of Agro-Environment Science, 2011.

[11] SMITH, S. V. Marine macrophytes as a global carbon sink.[J]. Science, 1981, 211(4484):838-840.

Author Index