Publisher: State and Provincial Joint Engineering Lab. of Advanced Network Monitoring and Control (ANMC)

Cooperate:

Xi'an Technological University (China)
West Virginia University (USA)
Huddersfield University of UK (UK)
Missouri Western State University (USA)
James Cook University of Australia (Australia)
National University of Singapore (Singapore)

Approval:

Library of Congress of the United States
Shaanxi provincial Bureau of press, Publication, Radio and Television

Address:

4525 Downs Drive, St. Joseph, MO64507 , USA
No. 2 Xuefu Road, weiyang District, Xi’an, 710021, China

Telephone: +1-816-2715618 (USA)   +86-29-86173290 (China)

Website：www.ijanmc.org

E-mail：ijanmc@jianmc.org
xxwlcn@163.com

ISSN: 2470-8038

Print No. (China): 61-94101

Editor in Chief

Associate Editor-in-Chief

Administrator

Associate Editors

Proceedings

2017 International Conference on Computer Network, Electronic and Automation

ICCNEA 2017

Xi'an, China
23-25 September 2017

CPS
Conference Publishing Services

The papers in this book comprise the proceedings of the meeting mentioned on the cover and title page. They reflect the authors' opinions and, in the interests of timely dissemination, are published as presented and without change. Their inclusion in this publication does not necessarily constitute endorsement by the editors, the IEEE Computer Society, or the Institute of Electrical and Electronics Engineers, Inc.

IEEE Computer Society Order Number E6156
ISBN-13: 978-1-5386-3981-8
BMS Part # CFP17M07-CDR

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/cspress
csbooks@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

Individual paper REPRINTS may be ordered at: <reprints@computer.org>

Editorial production by Juan E. Guerrero
Cover art production by Annie Jiu
Printed in the United States of America by Applied Digital Imaging

CPS
Conference Publishing Services

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

Organizing Committees

General Chair:
Ph.D. Weiguo Liu
Prof. and President of Xi'an Technological University, China
Director of State and Provincial Joint Engineering Lab. of Advanced Network and Monitoring, China

Chairman:
Ph.D. Houbing Song
Golden Bear Scholar and Professor, SM IEEE
Department of Electrical and Computer Engineering, West Virginia University, USA
Director of Security and Optimization for Networked Globe Laboratory, USA

Dr. Jianguo Wang
Prof. and Dean, Xi'an Technological University, China
Vice-Director of State and Provincial Joint Engineering Lab. of Advanced Network and Monitoring, China

Program Chair:
Prof. George Yang
Missouri Western State University, USA
4525 Downs Drive
St. Joseph, MO 64507
Phone: (816) 271-5618 (office)

Prof. Zhongsheng Wang
Xi'an Technological University, China
No. 4 North Jinhua Road,Xi'an 710032,Shaanxi Province,China
Tele:+86-29-86173290(office)

Table of Contents

2017 International Conference on Computer Network, Electronic and Automation (ICCNEA 2017)

Construction and Visualization of 3D Landscape

Pingping Liu
School of Computer Science and Engineering,Xi'an
Technological University,
Xi'an 710021,China
Email:1341369601@qq.com

Zhaopan Lu
School of Computer Science and Engineering,Xi'an
Technological University,
Xi'an 710021,China

Abstract—In order to know the operation condition before working, and reduce the risk of serious accidents of coal production industry. In this paper, by analyzing the characteristics of the environment of underground mine, using virtual reality technology, combined with the 3d Max modeling techniques and VC++ calling the OpenGL graphics interface techniques, the models are well optimized based on progressive mesh LOD algorithm for edge collapse. Using high-tech computer technology to simulate realistic mine production environment and realization of a friendly man-machine interactive virtual mine application system.

Keywords-Virtual Reality Technology; 3d Max; OpenGL; LOD algorithm

I. INTRODUCTION

With the development of virtual reality and related technology,3D digital geometric model has become a new generation of digital media types after digital audio, image and video, widely used in transportation, industrial manufacturing, military simulation etc. The optimization of 3D model is one of the key problems in the field of computer graphics and digital geometry processing. Based on the requirements of the application of digital mine geometry modeling, this paper aims at the problems such as the convenience of modeling, the diversity of modeling and the flexibility of classification, studying the modeling of digital mine and the organization of data and the simplification of data, a progressive mesh LOD algorithm based on edge collapse algorithm is proposed, which can optimize the virtual mine system model, improve the loading speed. Add the collision detection technology, collision detection technology, combined with 3d Max modeling technology and VC++ call OpenGL graphical interface technology to develop a set of digital mine application system, really and vividly reflect the mine operating conditions.

II. VIRTUAL REALITY TECHNOLOGY BASED ON MODELING

Virtual reality (VR) is a kind of advanced human computer interface, which can simulate and interact with many kinds of sensory channels, such as sight, hearing, touch, smell and taste. It makes full use of computer hardware and software technology, through the computer integrated technology to simulate the virtual three-dimensional space,providing a real-time virtual environment, with immersion and interaction and conception.

The core of the modeling method is to build the virtual model. The establishment of virtual model can use different modeling software, such as CAD, Autodesk Revit, 3d Max and so on. Taking into account the realistic degree of modeling and the realization of virtual mine application system on the VC++ platform and OpenGL interface, the 3d Max is used to build the model.

III. CONSTRUCTION AND VISUALIZATION OF 3D MODEL OF OPENGL

In the development of the system, through the combination of 3D Max software and OpenGL, we can reduce the complexity of the system. The key part of constructing the virtual mine model is the modeling of the underground scene. According to the state characteristics of the mine scene model, it can be divided into dynamic entity model and static entity model

A. *Construction of static entity model of virtual mine*

The mine static entity includes the mine shaft, the mine track in the mine, the mine pipeline, the circuit and the self rescue system and so on. Static modeling is mainly based on the physical characteristics of the physical model and the physical model of the physical environment. The physical modeling of the static entity model is mainly about the different solid textures caused by the external environment. The geometric model of the static entity model is the entity model of its own shape. In the process of geometric modeling of static solid model, depending on the complexity of the collected entity model data, for regular static entity model, 3D Max can be used to build the model, for irregular entity model, AutoCAD is used to modify the model outline, and then the model is established by 3D Max.

B. *Construction of dynamic entity model of virtual mine*

In the process of constructing mine system model, it is possible to change the internal structure of the original model when constructing the moving part, therefore, we should increase the degree of freedom of the link in the model file, set corresponding positioning coordinates, based on the degree of freedom, the motion of the model is analyzed to determine the kinematic relationship.

As an important part of virtual mine system model, the modeling process of dynamic entity model is:① Using 3D Max to construct the static entity model and dynamic entity model, in the process of building dynamic model, the degree of freedom increases; ②Enhanced dynamic entity model animation effect; ③The flash effect of dynamic model is presented in the model of virtual mine system

IV. OPTIMAL DESIGN OF 3D LANDSCAPE SYSTEM MODEL

Because of the complexity of the mine scene, when loading the model, complex model will cause the system CPU load is too large, the system will not work in coordination, the model output delay and picture frame refresh are too slow, which reduces the effect of the mine system. As shown in figure 1, for the two spheres of the same radius and the number of patches is not equal, the required system CPU load and memory capacity are different, as shown in Table 1.

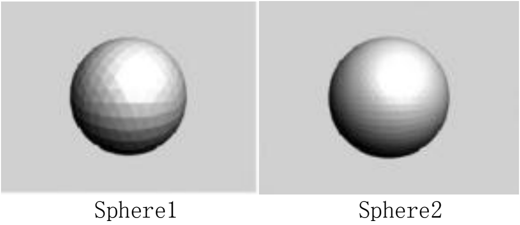


Figure 1. Different facets of sphere model

TABLE I. SYSTEM CPU LOAD COMPARE WITH MEMORY COMPARISON UNDER THE NUMBER OF DIFFERENT PIECES

Model name	Patch number	Occupied memory
Sphere 1	1000	18MB
Sphere 2	3500	28MB

In order to speed up the output of the scene model and ensure the real-time performance of the virtual scene, need to simplify the complexity of the scene. For the objects in the mine scene far away from the observer, there is no need to describe the details, the appropriate combination of some triangular surface of the object, the visual effect of the picture did not have an impact. The model details of virtual mine system are simplified by correlation, reduce the detail level of the model, After the treatment of the model can be selected according to specific occasions, and there is no need to select the full details of the model in any situation, and the number of triangles in scene model is greatly reduced. When the LOD algorithm is used to optimize the mine model, it is necessary to judge whether the scene objects need to be simplified and what level of detail should be used to represent the object, According to the characteristics of the complex structure of the virtual mine system, this paper proposes a progressive mesh LOD algorithm based on the edge collapse algorithm, the virtual mine system model is optimized by this algorithm.

A. *Optimization of progressive mesh LOD algorithm based on edge collapse*

1) The basic idea of progressive mesh

The basic idea of progressive mesh is the mesh simplification algorithm based on edge collapse, The triangular mesh edge collapse operation is shown in figure 2, a set of dual operations includes a side folding operation and a corresponding vertex splitting operation, Edge collapse operation is the edge (a, b) synthesis vertex {a},and delete the triangle {a, b, c} and {a, b, d};Contrary, Vertex splitting operations split vertex {a} into edges (a, b), and generate new triangles {a, b, c} and {a, b, d}.

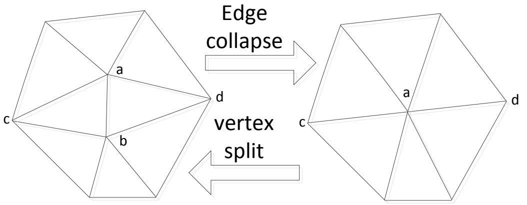


Figure 2. Sketch map of edge collapse operation

The feature of this method is that the local mesh elements are changed by each iteration, a vertex, an edge, and two triangles, therefore, it is possible to produce very regular information sequences. A set of edge collapse operation sequences will generate a set of continuous approximate grid sequences, can be expressed as

$$M_{\mathrm{n}} \xrightarrow{\mathrm{edgecol}_{\mathrm{n-1}}} M_{\mathrm{n-1}} \xrightarrow{\mathrm{edgecol}_{\mathrm{n-2}}} \cdots M_1 \xrightarrow{\mathrm{edgecol}_0} M_0 \qquad (1)$$

Vertex split conversion, through the simplest grid to get the subsequent high-level grid of detail, in theory, these two operations are reversible, can be expressed as

$$M_0 \xrightarrow{\mathrm{vsplit}_0} M_1 \xrightarrow{\mathrm{vsplit}_1} \cdots M_{\mathrm{n-1}} \xrightarrow{\mathrm{vsplit}_{\mathrm{n-1}}} M_{\mathrm{n}} \qquad (2)$$

2) Error measure of edge collapse

Error measure of edge collapse is the new vertex to a related set of plane and the square of the distance after edge collapse. This method is simple, small memory, and fast, the quality of the simplified mesh is very high. It is an ideal error measure which takes into account both the speed and the quality of the model.

In Euclidean space, the expression of the plane is

$$n^{\mathrm{T}} v_0 + d = 0 \qquad (3)$$

$n = [n_x n_y n_z]^{\mathrm{T}}$ is plane normal vector, is constant.

The square of the distance from point $v = [xyz]^{\mathrm{T}}$ to the plane can be expressed as

$$D^2 = (n^{\mathrm{T}} v + d)^2 = (v^{\mathrm{T}} n + d)(n^{\mathrm{T}} v + d) = v^{\mathrm{T}} (nn^{\mathrm{T}}) v + 2dn^{\mathrm{T}} v + d^2 \qquad (4)$$

Can define three tuples $Q = (a, b, c) = (nn^{\mathrm{T}}, dn, d^2)$, to express D^2 , Such as (5)

$$D^2 = Q(v) = v^{\mathrm{T}} av + 2b^{\mathrm{T}} v + c \qquad (5)$$

Where a is a symmetric matrix of 3*3, b is a 3 dimensional vector, c is a constant.

The 2 error measure can be used to facilitate error accumulation, such as formula (6)

$$Q_1(v) + Q_2(v) = (Q_1 + Q_2)(v) \qquad (6)$$

Among, $(Q_1 + Q_2) = (a_1 + a_2, b_1 + b_2, c_1 + c_2)$ 。

Sum the error matrix corresponding to each correlation plane of the vertex, The error matrix of the vertex is obtained.

In the grid, The sum of the error matrix of the two vertices of the edge of each edge is $Q = Q_1 + Q_2$,The two error value is $Q(v)$.

B. Implementation of the algorithm in virtual mine model optimization

Optimization model M_0 , M_1 , M_2 …need to maintain a certain degree of similarity with the original mesh model, the key is on the selection of edge collapse and the location of the target points after the edge collapse, These depend on the edge collapse cost. It is best to meet three conditions: ①Simple operation; ②Simplified record of original model to simplified model; ③Most mesh models can be simplified.

Based on the above criteria, in order to meet the actual needs of the mine, The effect of deleting a shorter edge on the overall model shape, the influence of a shorter edge on the shape of the whole model and the influence of the two error value on the flatness of the model are also discussed, this paper proposes a method to measure the edge collapse cost by using the product of side square and the two error value.

Edge collapse cost can be calculated as follows:

$$\mathrm{edgecolcost}(v)=Q(v)\times L^2 \tag{7}$$

In this type, $\mathrm{edgecol}\cos t(v)$ is edge collapse cost; $Q(v)$ two error value for the folded edge; L is the length of the side being folded.

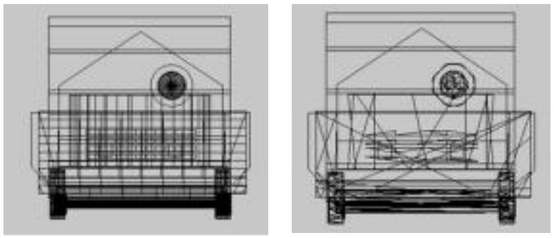
Selection of folded edges: The choice of each edge collapse should be folded at the cost of $\mathrm{edgecol}\cos t(v)$ from small to large order. In view of the special situation of the triangle shape distortion caused by the folding operation, So it is necessary to check the triangle mesh model which is generated after the edge collapse. If the new mesh model is reasonable, the edge collapse operation can be implemented, In the case of distortion and other special circumstances, the upper and lower edges of the set of edges are folded.

The location of the target point to be generated after the folded edge is selected. For large objects, a large amount of data for a wide range of 3D models, in order to improve the efficiency, the virtual mine model directly uses the original data. The original point is the folded edge vertex in a lower cost. Calculate the sizes of $\mathrm{edgecol}\cos t(v_1)$ and $\mathrm{edgecol}\cos t(v_2)$. Because the L is the same, in fact, compare the size of the $Q(v_1)$ and the $Q(v_2)$, select one of the smaller fold .

The algorithm is used to simplify a mine car in virtual mine, the data shown in Table 2. Figure 3 is the model grid of the mine, Among them (a) is the original details of the mine level diagram, and (b) is a rough layer model which is simplified by the algorithm. The result showed that the number of vertices and the number of triangles on the surface of the car are obviously reduced after simplification.

TABLE II. THE MODEL DATA OF DIFFERENT LEVELS OF CAR

cars model	Vertex number	Number of triangles
Before simplification	2672	5236
After simplification	413	718



(a) Fine mesh (b) Rough mesh

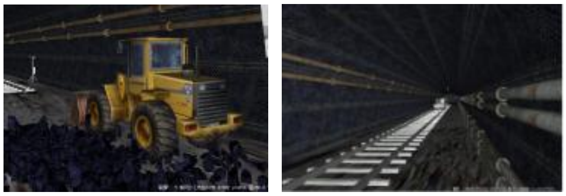
Figure 3. The mesh hierarchy of tramcar

V. REALIZATION OF VIRTUAL MINE 3D ROAMING

Virtual mine roaming system is through a camera to simulate the human eyes to observe the scene in the scene. In OpenGL, we mainly use gluLookAt (...) function to observe the virtual mine scene, which is used to change the position of the viewpoint in the Virtual Mine Scene. The position of this view represents the location of the user's eyes. When the user controls the main view roaming in the virtual mine scene, to see the distant scene model is more and more close, it shows that the position of the viewpoint in the virtual mine scene has changed. The gluLookAt（…）function is as follows:

gluLookAt(GLdoubleeyex,GLdoubleeyey,GLdoubleeyez,GLdoublecenterx,GLdoublecentery,GLdoublecenterz,GLdoubleupx,GLdoublupy,GLdoubleupz)

The gluLookAt（…）function is used to control the camera, the first three parameters indicate the location of the camera, the middle of the three parameters that represent the location of the point of view, the following three parameters represent the orientation of the camera, through the three sets of parameters to move the camera position, change the viewpoint to achieve camera roaming, The roaming effect is shown in Figure 4.



(a) mine shaftbottom (b) coal mine roadway

Figure 4. The effect chart of mine roaming

VI. CONCLUSION

1)Analysing the characteristics of the mine environment, the whole scene of the 3D virtual mine and the modeling of some equipment models are realized by using 3ds Max software. Combined with LOD algorithm, the progressive mesh simplification algorithm is improved. This paper proposes a progressive LOD algorithm based on edge collapse, which can realize the operation of edge collapse. It can meet the requirement of rapid generation of progressive mesh model, simple operation, fast loading, less memory.

2)Analysing the virtual mine system, a reasonable framework of the virtual mine system is established, which realizes the visualization of the 3D tunnel and the friendly interface, and completes the development of the virtual mine system.

SPONSORS OR SUPPORTERS

This word is partially supported by Special research project of Shaanxi Provincial Department of Education “16JK1376”.

AUTHORBRIEF

Liu Ping-Ping(1971-), female, Associate Professor, Xi'an Technological University, Research area: Artificial intelligence

REFERENCE

[1] CAI Zhibang.The Research on Virtual Reality System of Mine Safety[D].Henan: Henan Polytechnic University,2010,12.

[2] GUO Xiaohui,WANG Jing,YANG Yang,ZHANG Xin,XU Guanghua.Active and Passive Training System of Lower Limb Rehabilitation Based on Virtual Rerlity[J].Xian: JOURNAL OF XI’AN JIAOTONG UNIVERSITY,2016,50(2):124-131.

[3] YU Pei.The Research and Implementation of Virtual Reality System Based On Intelligent Video Technology[D].Chengdu: University of Electronic Science and Technology,2015,12.

[4] ZHAO Xin.RESEARCH AND DEVELOPMENT ON THE YELLOW RIVER MUSEUM BASE VIRTUAL REALITY WALKTHROUGH SYSTEM[D].Beijing:Beijing University of Technology,2015,4.

[5] SHUI Yong.Research on Continuous Collision Detection Algorithm In Virtual Reality[D].Anhui:University of Science and Technology of China, 2013,5.

Levenberg-Marquardt Method Based Iteration Square Root Cubature Kalman Filter ant its applications

Mu Jing
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, 710032, China;
e-mail:mujing1977@163.com

Wang Changyuan
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, 710032, China;
e-mail:cyw901@163.com

Abstract—**To improve the low state estimation accuracy of nonlinear state estimation due to large initial estimation error and nonlinearity of measurement equation, we obtain Levenberg-Marquardt (abbr. L-M) method based iteration square root cubature Kalman filter (ISRCKFLM) combining the measurement update of square root cubature Kalman filter (SRCKF) with nonlinear least square error, so the ISRCKFLM algorithm has the virtues of global convergence and numerical stability. We apply the ISRCKFLM algorithm to state estimation for re-entry ballistic target; the simulation results demonstrate the ISRCKFLM algorithm has better accuracy of state estimation.**

Keywords-Nonlinear filtering; Cubature Kalman filter; Levenberg-Marquardt method

I. INTRODUCTION

A series of nonlinear filters have been developed to apply to state estimation for the last decades. Up to now the commonly used non-linear filtering is the extended Kalman filter (EKF) [1, 2]. The EKF is based on first-order Taylor approximations of state transition and observation equation about the estimated state trajectory under Gaussian assumption, so EKF may introduce significant bias, or even convergence problems due to the overly crude approximation [3].

Recently, one type of suboptimal nonlinear filters based on numerical multi-dimensional integral were introduced in [4-6], such as cubature rules based cubature Kalman filter (CKF) and the interpolatory cubature Kalman filters (ICKFs), which used numerical multi-dimensional integral to approximate the recursive Bayesian estimation integrals under the Gaussian assumption. The CKF can solve high-dimensional nonlinear filtering problems with minimal computational effort and can be deemed as special case of ICKFs. Furthermore, the stability of CKF for non-linear systems with linear measurement is analyzed and the certain conditions to ensure that the estimation error of the CKF remains bounded are proved in [7]. On the other hand, in order to decrease the effect of initial estimation error and nonlinearity of measurement equation, Levenberg-Marquardt method based iteration cubature Kalman filter was developed on the basis of the CKF in Reference [8]. In fact, singular matrix occurs in the implementation of the above filters mentioned if the initial estimation is selected improperly. So the cubature rule is exploited as square root cubature information filter [9] and the square root cubature Kalman filter (SRCKF) was developed in order to mitigate ill effects and improve the numerical stability [5].The SRCKF also shows its weakness in the robustness and estimation accuracy. Making use of L-M method and the superiority of the SRCKF algorithm, we obtain the L-M method based iterative square root cubature Kalman filter (ISRCKFLM), in which, we transform the measurement update of SRCKF to the problem of nonlinear least square error, then use L-M method to solve it and obtain the optimal state estimation and covariance to improve the low state estimation accuracy of nonlinear state estimation due to large initial estimation error and nonlinearity of measurement equation.

The rest of the paper is organized as follows. We begin in Section 2 with a description of square root cubature Kalman filter (SRCKF). The L-M method based iterative square root cubature Kalman filter (ISRCKFLM) is developed in Section 3. Then we apply the ISRCKFLM algorithm to track re-entry ballistic target (RBT) with unknown ballistic coefficient and discuss the simulation results in Section 4. Finally, Section 5 concludes the paper.

II. L-M BASED ITERATION SQUARE ROOT CUBATURE KALMAN FILTER

Consider the following nonlinear dynamics system:

$$\boldsymbol{x}_k = \boldsymbol{f}(\boldsymbol{x}_{k-1}) + \boldsymbol{w}_{k-1}\,. \tag{1}$$

$$\boldsymbol{z}_k = \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{v}_k\,. \tag{2}$$

where $\boldsymbol{f}$ and $\boldsymbol{h}$ are some known nonlinear functions; $\boldsymbol{x}_k \in \mathbb{R}^{n_x}$ and $\boldsymbol{z}_k \in \mathbb{R}^{n_z}$ is state and the measurement vector, respectively; $\boldsymbol{w}_{k-1}$ and $\boldsymbol{v}_k$ are process and measurement Gaussian noise sequences with zero means and covariance $\boldsymbol{Q}_{k-1}$ and $\boldsymbol{R}_k$, respectively, and $\{\boldsymbol{w}_{k-1}\}$ and $\{\boldsymbol{v}_k\}$ are mutually uncorrelated.

Suppose that the state distribution at k-1 time is $\mathbf{x}_{k-1} : \mathrm{N}(\hat{\mathbf{x}}_{k-1}, \mathbf{S}_{k-1}\mathbf{S}_{k-1}^T)$, Levenberg-Marquardt based Iteration square root cubature Kalman filter (ISRCKFLM) is described as follows.

(1) Time Update

1) Calculate the cubature points and propagate the cubature points through the state equation

$$\boldsymbol{X}_{i,k-1} = \boldsymbol{S}_{k-1}\xi_i + \hat{\boldsymbol{x}}_{k-1} . \tag{3}$$

$$\boldsymbol{X}_{i,k}^{*} = \boldsymbol{f}(\boldsymbol{X}_{i,k-1}) . \tag{4}$$

where $\xi_i = \sqrt{m/2}[1]_i, \omega_i = 1/m, i = 1, \cdots m = 2n_x$, the $[1]_i$ is a n_x dimensional vector and is generated according to the way described in [2].

2) Evaluate the predicted state and square root of the predicted covariance

$$\bar{\boldsymbol{x}}_k = \sum_{i=1}^{m} \omega_i \boldsymbol{X}_{i,k}^{*} . \tag{5}$$

$$\bar{\boldsymbol{S}}_k = Tria([\boldsymbol{\chi}_k^{*} \;\; \boldsymbol{S}_{Q,k-1}]) . \tag{6}$$

here, $\boldsymbol{S}_{Q,k-1}$ denotes a square-root factor of $\boldsymbol{Q}_{k-1}$ and *Tria*() is denoted as a general triagularization algorithm. The matrix $\boldsymbol{\chi}_k^{*}$ is defined as:

$$\boldsymbol{\chi}_k^{*} = 1/\sqrt{m}[\boldsymbol{X}_{1,k}^{*} - \bar{\boldsymbol{x}}_k \;\; \boldsymbol{X}_{2,k}^{*} - \bar{\boldsymbol{x}}_k, \cdots, \boldsymbol{X}_{m,k}^{*} - \bar{\boldsymbol{x}}_k] . \tag{7}$$

3) Evaluate the modified covariance:

$$\tilde{\boldsymbol{P}}_k = \left[\boldsymbol{I} - \bar{\boldsymbol{S}}_k \bar{\boldsymbol{S}}_k^T \left(\bar{\boldsymbol{S}}_k \bar{\boldsymbol{S}}_k^T + \frac{1}{\mu_i} \boldsymbol{I} \right)^{-1} \right] \bar{\boldsymbol{S}}_k \bar{\boldsymbol{S}}_k^T . \tag{8}$$

where is adjusting parameter.

(2) Measurement update

1) Set the initial value as: $\hat{\boldsymbol{x}}_k^{(0)} = \bar{\boldsymbol{x}}_k$.

2) Assuming the *i*-th iterate $\hat{\boldsymbol{x}}_k^{(i)}$, calculate the matrix

$$\boldsymbol{L}_k^{(i)} = \tilde{\boldsymbol{P}}_k \boldsymbol{J}_h^T(\hat{\boldsymbol{x}}_k^{(i)}) \left[\boldsymbol{J}_h(\hat{\boldsymbol{x}}_k^{(i)}) \tilde{\boldsymbol{P}}_k \boldsymbol{J}_h^T(\hat{\boldsymbol{x}}_k^{(i)}) + \boldsymbol{R}_k \right]^{-1} . \tag{9}$$

3) Calculate the *i*-th iterate

$$\begin{aligned} \hat{\boldsymbol{x}}_k^{(i+1)} &= \bar{\boldsymbol{x}}_k + \boldsymbol{L}_k^{(i)} \left\{ \boldsymbol{z}_k - \boldsymbol{h}(\boldsymbol{x}_k^{(i)}) - \boldsymbol{J}_h(\hat{\boldsymbol{x}}_k^{(i)})(\bar{\boldsymbol{x}}_k - \hat{\boldsymbol{x}}_k^{(i)}) \right\} \\ &\quad - \mu_i \left\{ \boldsymbol{I} - \boldsymbol{L}_k^{(i)} \boldsymbol{J}_h(\hat{\boldsymbol{x}}_k^{(i)}) \right\} \tilde{\boldsymbol{P}}_k (\bar{\boldsymbol{x}}_k - \hat{\boldsymbol{x}}_k^{(i)}) \end{aligned} . \tag{10}$$

4) Calculate the iteration termination condition

$$\left\| \hat{\boldsymbol{x}}_k^{(i+1)} - \hat{\boldsymbol{x}}_k^{(i)} \right\| \le \varepsilon \text{ ορ } i = N_{\max} . \tag{11}$$

ε and $N_{\max}$ are predetermined threshold and maximum iterate number, respectively. If the termination condition meets, the iterate return to 5); otherwise, set $\hat{\boldsymbol{x}}_k^{(i)} = \hat{\boldsymbol{x}}_k^{(i+1)}$, continue to 2).

5) Calculate the state estimation at k time instant

$$\hat{\boldsymbol{x}}_k = \hat{\boldsymbol{x}}_k^{(N)} . \tag{12}$$

6) Evaluate the cross-covariance and square root of innovation covariance at k time

$$\boldsymbol{P}_{xz} = \bar{\boldsymbol{S}}_k \bar{\boldsymbol{S}}_k^T \boldsymbol{J}_h^T(\hat{\boldsymbol{x}}_k^{(N)}) . \tag{13}$$

$$\boldsymbol{S}_{zz} = Chol(\left[\boldsymbol{J}_h(\hat{\boldsymbol{x}}_k^{(N)}) \bar{\boldsymbol{S}}_k \quad \boldsymbol{S}_{R,k} \right]) . \tag{14}$$

7) Calculate the square root of covariance at *k* time

$$\boldsymbol{K}_k = \boldsymbol{P}_{xz} / \boldsymbol{S}_{zz}^T / \boldsymbol{S}_{zz} . \tag{15}$$

$$\boldsymbol{S}_k = Chol(\left[\bar{\boldsymbol{S}}_k - \boldsymbol{K}_k \boldsymbol{J}_h(\hat{\boldsymbol{x}}_k^{(N)}) \bar{\boldsymbol{S}}_k \quad \boldsymbol{K}_k \boldsymbol{S}_{R,k} \right]) . \tag{16}$$

where symbol “/ ” represents the matrix right divide operator.

III. Applications to State Estimation for Re-entry Ballistic Target

To demonstrate the performance of the ISRCKFLM algorithm, we apply the ISRCKFLM to estimate state of re-entry ballistic target with unknown ballistic coefficient and compare its performance against the SRCKF and iterate square root cubature Kalman filter using Gauss-Newton method (ISRCKF) algorithms. All the simulations were done in MATLAB on a ThinkPad PC with an Intel (R) CORE i5 M480 processor with the 2.67GHz clock speed and 3GB physical memory.

In the simulation, the parameters and the initial state estimate are the same as in [10]. To demonstrate the performance of the ISRCKFLM algorithm, we use the root-mean square error (RMSE) and average accumulated mean-square root error (AMSRE) in the position, velocity and ballistic coefficient introduced in [8]. Figure. 1, Figure. 2 and Figure. 3 show the RMSEs for the SRCKF, ISRCKF and ISRCKFLM (μ=10^{-10}) in position, velocity and ballistic coefficient in an interval of 15s-58s. The AMSREs of the three filters in position, velocity and ballistic coefficient are listed in Table. 1. The iteration number selected in the ISRCKFLM and ISRCKF algorithms is 4. All performance curves and figures in this subsection were obtained by averaging over 100 independent Monte Carlo runs. All the filters are initialized with the same condition in each run.

From Figure. 1, we can see that the RMSE of ISRCKFLM in position is far less than that of SRCKF algorithm, and is less than that of ISRCKF algorithm. Moreover, the ISRCKFLM needs 14.5 seconds to make the RMSE in position reduce below 500 meters, the ISRCKF algorithm needs 34.6 seconds, and SRCKF algorithm needs about 47.6 seconds, so the ISRCKFLM algorithm has faster convergence rate than the SRCKF and ISRCKF algorithms. So the estimates provided by the ISRCKFLM in the position and velocity are markedly better than those of SRCKF and ISRCKF algorithms.

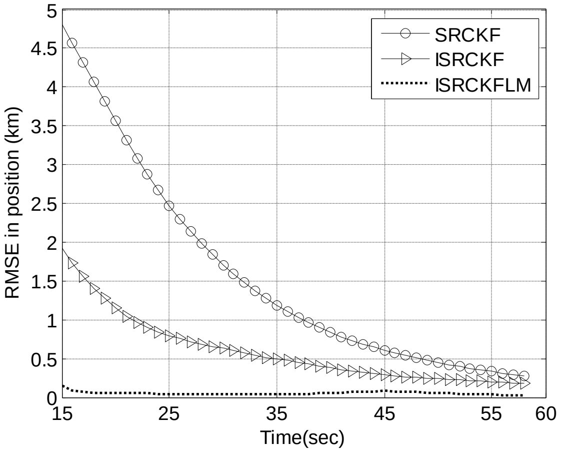


Figure 1. RMSEs in position for various filters

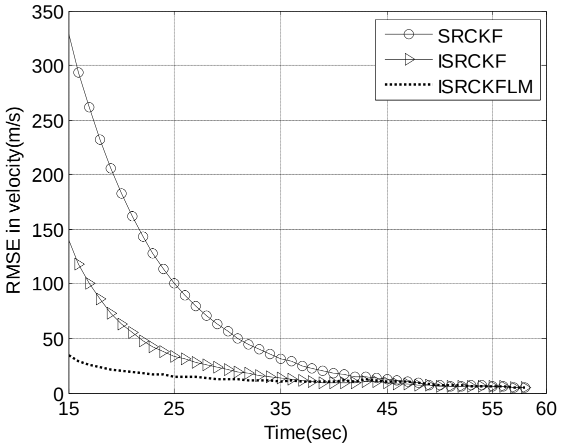


Figure 2. RMSEs in velocity for various filters

Observe from Figure. 2, the RMSE of ISRCKFLM in velocity is far less than those of SRCKF and ISRCKF algorithm in the interval time (t < 35s), the ISRCKFLM still has faster convergence rate. And the RMSEs of the three filters lie at the lower level in the period (t >35s).

As to the estimation of the ballistic coefficient, in the Figure. 3, the RMSEs of the three filters have less improvement in the interval time (0< t <35s) because of having less effective information about it from the noisy measurement. The RMSEs of the three filters begin to decrease at about t=37s because the measurements have the effective information on ballistic coefficient. In the period (35s< t <45s), the RMSE of the ISRCKFLM algorithm for the ballistic coefficient decreases more rapidly than that of SRCKF, and decreases at the same rate as that of ISRCKF. At the period 45s< t < 58s, the RMSE in the ISRCKFLM algorithm decreases most rapidly among the three algorithms. The ballistic coefficient estimate in the ISRCKFLM algorithm has the great improvement.

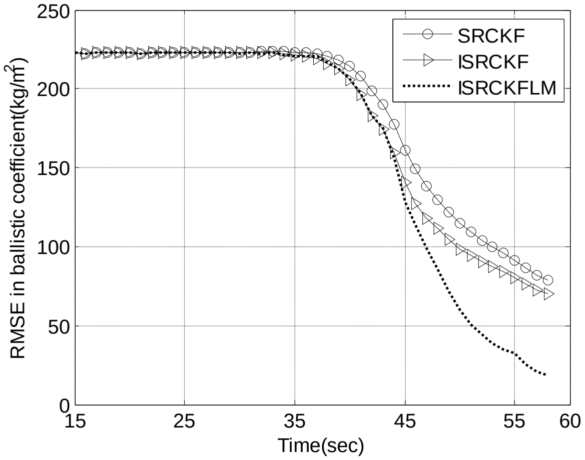


Figure 3. RMSEs in ballistic coefficient for various filters

TABLE I. AMSREs IN POSITION, VELOCITY AND BALLISTIC COEFFICIENT

Algorithms	AMSREp (m)	AMSREv (m/s)	AMSRE$_{\beta}$ (kg/m^2)
SRCKF	2693.096	306.133	165.363
ISRCKF	1457.078	250.900	162.530
ISRCKFLM	856.993	220.296	160.658

According to Figure.1-Figure. 3, the RMSEs of ISRCKFLM in position and velocity markedly decrease, compared with those of the SRCKF and ISRCKF algorithm. Although the RMSE of ISRCKFLM in ballistic coefficient has less improvement, its RMSE significantly reduces in the last period. So the ISRCKFLM improves the state estimation accuracy of re-entry ballistic target.

From TABLE 1. 1, it is seen that, the ISRCKFLM's AMSRE in position reduces by about 68%, and its AMSRE in velocity reduces by about 28% compared to SRCKF. And compared to ISRCKF, the AMSRE of ISRCKFLM algorithm in position decreases by about 41%, and its AMSRE in velocity decreases by about 12%. Table.1 shows ISRCKFLM's AMSRE in ballistic coefficient reduces marginally, but Figure.3 shows the ISRCKFLM's RMSE is less than the other two filters in the interval of 40s-58s. Hence, the ISRCKFLM is to be preferred over the other filters in the light of AMSREs in the position, velocity and ballistic coefficient and has better performance.

Therefore, on the basis of the simulation results presented in Figure.1-Figure.3 and Table.1, one can draw a conclusion that the ISRCKFLM algorithm yields on the superior performance over the SRCKF and ISRCKF algorithms on state estimation of re-entry ballistic target.

IV. CONCLUSION

The ISRCKFLM algorithm has the advantages of global convergence, fast convergence and numerical stability. The

ISRCKFLM algorithm is applied to state estimation for re-entry ballistic target. Simulation results demonstrate that the performance of ISRCKFLM algorithm is superior to SRCKF and ISRCKF algorithms. So the ISRCKFLM algorithm is much more effective and improves the performance of state estimation to a marked degree.

ACKNOWLEDGMENT

The authors would like to thank the support of the State and Local Joint Laboratory of Advanced Network and Monitoring Control Engineering (No. GSYSJ201603).

REFERENCES

[1] Bar-Shalom, Y.; Li, X.R. Kirubarajan, T. Estimation with Applications to Tracking and Navigation. New York: John Wiley &Son, 2001.

[2] Grewal, M.S.; Andrews, A.P. Kalman filtering: theory and practice using Matlab. New York: John Wiley & Sons, 2008.

[3] Julier, S. J.; Uhlmann, J.K, “Unscented filtering and nonlinear estimation,” Proceedings of IEEE, 2004, vol.92,issue.12, 2004, pp. 401-422.

[4] Arasaratnam, I.; Haykin, S.; Hurd, T.R, “Cubature Kalman filtering for continuous-discrete systems: theory and simulations,” IEEE Transaction on Signal Processing, vol.58, issue10, 2010, pp. 4977-4993.

[5] Arasaratnam, I.; Haykin, S, “Cubature Kalman filters,” IEEE Transactions on. Automatic Control., vol.54, issue 6, 2009, pp.1254-1269.

[6] Zhang, Y.G.; Huang, Y. L.; Li, N.; Zhao, L, “Interpolatory cubature Kalman filters,” IET Control Theory & Applicaitons. Vol.9, issue 11, 2015, pp.1731–1739.

[7] Jafar, Z.; Ehsan, S, “Convergence analysis of non-linear filtering based on cubature Kalman filter,” IET Science Measurement & Technology. Vol.9, issue 3, 2015, pp.294–305.

[8] Mu Jing, Cai Yuanli, Wang Changyuan, “L-M Method Based Iteration Cubature Kalman Filter and Its Applications,” Journal of Xi’an Technoligical University, vol.33, issue 1, 2013, pp.1-6.

[9] Chandra, K.P.B.; Da-Wei, G.; Postlethwaite, I, “Square root cubature information filter,” IEEE Sensors Journal, vol. 13, issue 2, 2013, pp.750–758.

[10] Mu. J.; Cai. Y. “Likelihood-based iteration square-root cubature Kalman filter with applicaitons to state estimation of re-entry ballistic target,” Transactions of the Institute of Measurement and Control., vol.35, issue 7, 2013, pp. 949-958.

Multi - scale Target Tracking Algorithm with Kalman Filter in Compression Sensing

Yichen Duan
Department of Electronic and Information Engineering
Xi`an China
13488232750@163.com,wp_xatu@163.com

Xue Li
Department of Electronic and Information Engineering
Xi`an China
13488232750@163.com,wp_xatu@163.com

Peng Wang
Department of Electronic and Information Engineering
Xi`an China
13488232750@163.com,wp_xatu@163.com

Dan Xu
Department of Electronic and Information Engineering
Xi`an China
13488232750@163.com,wp_xatu@163.com

***Abstract*—Real-time Compressive Tracking (CT) uses the compression sensing theory to provide a new research direction for the target tracking field. The algorithm is simple, efficient and real-time. But there are still shortcomings: tracking results prone to drift phenomenon, cannot adapt to tracking the target scale changes. In order to solve these problems, this paper proposes to use the Kalman filter to generate the distance weights, and then use the weighted Bayesian classifier to correct the tracking position, and perform multi-scale template acquisition in the determined position to adapt to the changes of the target scale. Finally, introducing the adaptive learning rate while updating to improve the tracking effect.. Experiments show that the improved algorithm has better robustness than the original algorithm on the basis of maintaining the original algorithm real-time.**

Keywords-compression sensing; CT; multi-scale; Kalman filter

I. INTRODUCTION

Target tracking is the core research content in the field of machine vision. It has a wide range of applications in human-computer interaction, video surveillance, scene comprehension and behavior recognition. In recent years, domestic and foreign scholars have proposed a variety of tracking methods, such as target-based, regional matching tracking algorithm, but these algorithms in the practical application of the situation in the poor robust, easy to track failure. Most of the current tracking problems using background and target binary classification of ideas, that is tracking-by-detection.

Compressed sensing theory is a kind of signal expression based on sparse expression in recent years, which has great influence in mathematics and engineering application. Has been applied to wireless communications, image processing, pattern recognition in many areas. Kaihua Zhang et al[8]. Applied the compression perceptual theory to the target tracking problem. The algorithm is simple and efficient, and real-time，which provides a new research direction for the target tracking field, but the algorithm still has some shortcomings .

Compression sensing algorithm in the tracking process tracking frame scale unchanged, easy to introduce background error, resulting in tracking box drift, the final tracking failure. Aiming at this problem, a simple scale transformation method is proposed, which can adapt to the real-time performance at the same time. Compression tracking algorithm is a typical tracking-by-detection framework, once the target part of the block, must introduce the error, resulting in drift, the location is difficult to restore. In response to this problem, this paper proposes the use of Kalman filter tracking box position correction. Compression tracking classifier update using fixed update mode, that is, a fixed learning rate, this approach will inevitably lead to update rate cannot adapt to the target changes, at the same time easy to introduce errors, and then drift. This paper presents a relatively simple way to update, can be better adaptive target changes.

II. REAL-TIME COMPRESSIVE TRACKING

The compression perceptual tracking algorithm is a tracking algorithm based on the compression perceptual theory [1]. The compression perceptual theory[2] states that if a signal can be compressed and the random perceptual matrix satisfies the Johnson-Lindenstrauss inference[3], there is a higher probability that the Y is reconstructed by X .

$$V = \phi X \ \ \phi \in R^{m \times n}(m << n) \tag{1}$$

The compression tracking uses the formula (1) to extract the target feature[4], X is the characteristic matrix of the target high dimension, ϕ is the random perception matrix, V is the low-dimensional characteristic matrix of the target, and its sparse Chengdu depends on the sparse degree of ϕ . The compression tracking uses a very sparse and satisfying Johnson-Lindenstrauss inference of a random projection matrix defined as follows:

$$v_{ij} = \sqrt{s} \times \begin{cases} 1 & with \quad prob \quad \frac{1}{2s} \\ 0 & with \quad prob \quad 1-\frac{1}{s} \\ -1 & with \quad prob \quad \frac{1}{2s} \end{cases} \tag{2}$$

In the formula, s randomly selected from 2 to 4. Compression feature extraction target tracking algorithm is used with similar Hear-Like relative difference feature. Each element of the low-dimensional feature is a linear combination of spatial distributions of different scales.

After the compression feature is extracted, the compressed features are entered into the naive Bayesian classifier to distinguish the background and the target. The target low dimension represents $V(z)=(v_1,...,v_n)^T \in R^n$, assuming that each element is independent of each other, naive Bayesian model:

$$H(v)=\log\left[\frac{\prod_{i=1}^{n} p(v_i \mid y=1)p(y=1)}{\prod_{i=1}^{n} p(v_i \mid y=0)p(y=0)}\right]=\sum_{i}^{n}\log\left(\frac{p(v_i \mid y=1)}{p(v_i \mid y=0)}\right) \tag{3}$$

Provable high-dimensional random vector mathematically random projection is always in line with the Gaussian distribution [5]. Thus the conditional distributions $p(v_i \mid y=1)$ and $p(v_i \mid y=0)$ in the classifier conform to the Gaussian distribution.

$$p(v_i \mid y=1) \sim N(\mu_i^1, \sigma_i^1) \tag{4}$$

$$p(v_i \mid y=0) \sim N(\mu_i^0, \sigma_i^0) \tag{5}$$

In the formula, μ_i^1 and σ_i^1 represent the mean and standard deviation of the $i-th$ feature of the positive sample, respectively. μ_i^0 and σ_i^0 are the mean and standard deviation of the $i-th$ feature of negative samples, respectively. The maximum response value position is the most likely target location. After determining the maximum corresponding position, the relevant parameters can be updated from the adaptation background and target changes. Parameter updating formula is:

$$\mu_i^1 \longleftarrow \lambda\mu_i^1 + (1-\lambda)\mu^1 \tag{6}$$

$$\sigma_i^1 \longleftarrow \sqrt{\lambda(\sigma_i^1)^2 + (1-\lambda)(\sigma^1)^2 + \lambda(1-\lambda)(\mu_i^1-\mu^1)^2} \tag{7}$$

In the formula, λ is the learning rate , $\lambda > 0$ and λ is constant, the value reflects the speed of the updating. μ^0 and σ^0 update the same as above.

III. Improvement of Compression Tracking

A. Combined with Kalman filter correction

Compression tracking classifier formula (3) to determine the target area is the way to all the probability of a simple sum, the response to the maximum value of the region is the target location. But sometimes the characteristics of the target and background characteristics are similar, it is not conducive to the classification of the classifier. Assuming the true position of each frame, it is clear that the closer to the real position, the greater the probability. At this point can be generated distance weight, add it to the classifier, improve the classifier classification performance, to enhance the credibility of the classifier. Of course, the real location of the target cannot be known in advance, you can use the forecast position instead of the real location. Maintaining the Integrity of the Specifications.

Kalman filter is a linear system state equation, through to the system input and output data, the optimal estimation of the system state is presented. The Kalman filter can be used to estimate the target position of the next frame using the current motion target parameters [7].

$$X(k \mid k-1) = AX(k-1 \mid k-1) + BU(k) \tag{8}$$

In the formula, $X(k \mid k-1)$ is the upper frame target motion parameter. $X(k-1 \mid k-1)$ is the target tracking result for the last frame. $U(k)$ is the control amount, set to 0.

$$P(k \mid k-1) = AP(k-1 \mid k-1)A + Q \tag{9}$$

In the formula, $P(k \mid k-1)$ is the covariance corresponding to $X(k \mid k-1)$, and Q is the covariance of the target tracking system. The Kalman filter combines the predicted position of the current target with the current target. Calculate the optimal estimate of the position of the target $X(k \mid k)$. $K_g(k)$ is Kalman gain.

$$X(k \mid k) = X(k \mid k-1) + K_g(k)(Z(k) - HX(k \mid k-1)) \tag{10}$$

After the predicted position obtained by the Kalman filter, the distance of each sample is measured to obtain the position weight, and the position weight is added to the classifier. Define the distance from the sample to the target location:

$$l_i = \sqrt{\left[o_{\det}^k(x) - o_{pre}(x)\right]^2 + \left[o_{\det}^k(y) - o_{pre}(y)\right]^2} \tag{11}$$

In the formula, $o_{pre}(x)$ and $o_{pre}(y)$ predict the x coordinate and y coordinate of the target position by Kalman filter respectively. $o_{\det}^k(x)$ and $o_{\det}^k(y)$ are the x coordinates and y coordinates of the $i-th$ sample, respectively. In order to reduce the impact of noise on the weight of the reference[6], the normalization function is a hyperbolic tangent function. The normalized position weights are:

$$w_i = \frac{1}{2}\left\{\tanh\left[0.01 \times \left(\frac{\frac{1}{l_i} - \mu}{\sigma}\right) + 1\right]\right\} \tag{12}$$

In the formula, $1/l_i$ indicates that the distance from the predicted distance of the sample is, the smaller the probability that the sample is judged as the target, the smaller the position weight w_i is. When the sample distance is closer to the predicted position, the opposite is true. μ and σ are the mean and variance of $1/l_i$ respectively. The position weight is introduced into the formula (3):

$$H^{'}(v) = \log\left[\frac{\prod_{i=1}^{n} w_i p(v_i \mid y=1) p(y=1)}{\prod_{i=1}^{n} (1-w_i) p(v_i \mid y=0) p(y=0)}\right] = \sum_{i}^{n} \log\left(\frac{w_i\ p(v_i \mid y=1)}{(1-w_i) p(v_i \mid y=0)}\right) \tag{13}$$

By the normalization of $\tanh$, when w_i is 0.5, it does not affect the classifier. w_i greater than 0.5 when the $p(v_i \mid y=1)$ value becomes larger, less than 0.5 on the contrary. That is, when the predicted position is closer to the target, the W value is larger, and the closer the target response value is, the better the background and the target will be. This weighted Bayesian classifier improves the performance of the classifier and enhances the reliability of the classifier.

B. *Scale Processing*

The compression algorithm is fixed in scale, i.e., the size of the tracking window is constant. The theoretical analysis shows that when the target becomes larger, the negative sample acquisition area and the target area are too close, so that the generated classifier can reduce the performance of the target background. When the target becomes smaller, it is easy to introduce the background error, resulting in tracking box drift.

When the target position is obtained, it is assumed that the position is accurate and the scale is suitable and no error is introduced. The secondary acquisition can be carried out at this position. The acquisition is mainly carried out at this position for multiple scale templates, and the classifier is used again. Get the maximum response value, that is, the result of scale transformation.

This method assumes that the target does not undergo dramatic changes in the scale during the tracking process. So the collection of different scales of the size of the rectangular box can be in the up and down about four directions can be outward to the outside of the proportion of amplification and reduction. In this paper, the scale is chosen to be between 0.5 and 1.5, with an interval of 0.05. That is, you can collect a total of 20 different sizes of templates.

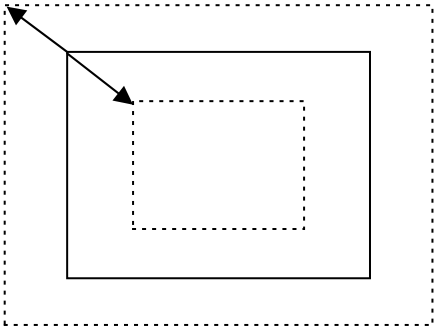


Figure 1. Scale Change

After the tracking frame scale changes, high-dimensional feature vectors cannot be mapped into low-dimensional space. In order to solve this problem, the essence of the nonzero term in the random sparse matrix is to sample the pixels in the tracking frame. After scale transformation, it is necessary to ensure that the position of the pixel sample and the relative position of the tracking frame remain unchanged. In practice, the random sparse matrix not only records the randomly generated weights, but also records the position of the feature expression. The operation of the feature is similar to that of the tracking frame, and the ratio of the scale and the scale of the tracking frame is the same in the four directions of the upper, lower, left and right directions, and the new random sparse matrix is obtained.

The above analysis shows that the new tracking box scale transformation method is simple. The new random sparse matrix is linearly transformed on the original random sparse matrix, and the overall computation is not large, and the real-time effect of the algorithm is limited. The scale transformation of this method can adapt to the change of target scale on the basis of keeping the algorithm real-time.

C. *Update Improvement*

From the perspective of algorithm design, the algorithm needs to be able to adapt to the appearance of the target to a certain extent. As can be seen from the above algorithm, the compression sensing algorithm will re-acquire the positive and negative samples and update the classifier after finding the "target position" of the frame. This approach is, to some extent, the appearance of the adaptive tracking target.

However, once the target is partially blocked, the background is complex, it will inevitably introduce noise and background error. In the follow-up of the process of tracking the phenomenon of drift, with the drift of the phenomenon of accumulation, and ultimately tracking failure.

Equation (6) (7) analysis shows that the new parameter model has two parts. The first part is the model of the previous frame, which represents the stability of the target. The second part is based on the current frame of the target for the collection of positive and negative samples to learn the new model, represents the goal of change. The new and old models are linearly combined to form an updated classifier parameter model. The learning rate of the compression perceptual tracking algorithm is a fixed value. When the new template is suspicious, it cannot effectively suppress the update.

This paper proposes a learning rate that can be adapted to a target change. The histogram is calculated for the maximum response position of the classifier, and the distance between the previous frame and the current frame processing result is calculated using the Bhattacharyya distance.

$$L = \sum_{i=1}^{n} \sqrt{p_i q_i} \quad L \in (0,1] \qquad (14)$$

The closer the distance, the higher the matching degree of the two images, on the contrary, the lower. Set a threshold at that time, you can update, on the contrary, it means that the current frame and the previous frame of the image difference is large, then refused to update to avoid the introduction of error. The new learning rate is:

$$\lambda^{'} = \lambda / L \qquad (15)$$

In the formula, A is the learning rate, and B is the Bhattacharyya distance of the previous frame and the current frame processing result. The larger the results are similar, the smaller the learning rate is needed, the smaller the difference is, the larger the learning rate is. In this way, adaptive control classifier learning rate. Because only need to calculate the previous frame and the next frame of the results of the histogram, and then Bhattacharyya distance operation, the overall calculation of the amount did not significantly improve, in keeping the algorithm on the basis of real-time better adapt to the appearance of the object changes.

D. *Algorithm*

```
Multi - scale Target Tracking Algorithm with Kalman Filter in Compression Sensing

Input: the $t-th$ image frame
1: Collect the image set for the first time and use the classifier to determine the target position.
2: The Kalman filter corrects the target position.
3: Scale transformation, and change the number of random sparse matrix columns, to ensure that the high-dimensional feature vector mapping to low-dimensional space.
4: The second collection of image sets, and extract low-dimensional features, update the classifier parameters.
Output: Tracking target location
```

IV. EXPERIMENT AND CONCLUSION ANALYSIS

A. *Tracking Effect Analysis*

This algorithm simulation platform for Visual Studio 2013, release mode, call opencv visual library programming, version number: 2.4.13. Tests use the same computer, and the test sequence comes from the common test set.

In this paper, the first group of test sequence name box, the background is complex, the target part of the occlusion and appearance changes. At the time of sequence 321, when the tracking target is partially blocked, the compression tracking algorithm begins to drift, and the 339th frame compression tracking algorithm fails to track completely. And the use of this algorithm, 324 objects occlusion can also be more accurate tracking to the target, in the first 339 series can continue to track the target. When the algorithm is improved, the tracking failure caused by the occlusion of the object is greatly reduced.

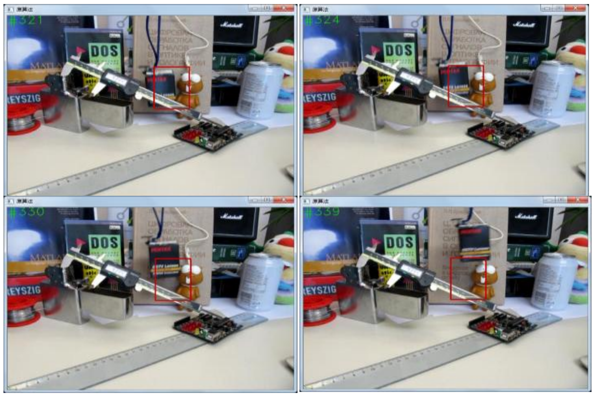


Figure 2. Box sequence of the original algorithm 321,324,330,339 frames

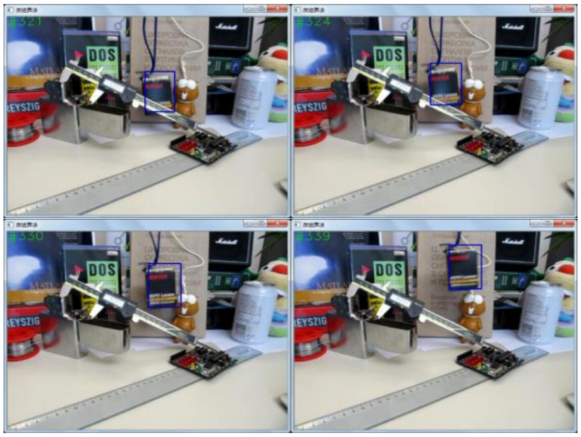


Figure 3. Box sequence of the improve algorithm 321,324,330,339 frames

In this paper, the second group of test sequence name Walking2, the background is more single, the target part of the block, in the tracking process has a similar target interference. In the 29th frame of the sequence, similar objects appear at 210 frames, and the similar target is partially blocked when the target is about 210 frames. At this time, the original compression tracking algorithm starts to track and drift, 220 frames are completely disturbed by similar target, 230 frame tracking algorithm is tracked failure. The use of this algorithm, in the 210 frame when the target is partially blocked in the case, to continue to track, 230 can also continue to track. After the algorithm is improved, the robustness of similar interference is improved.

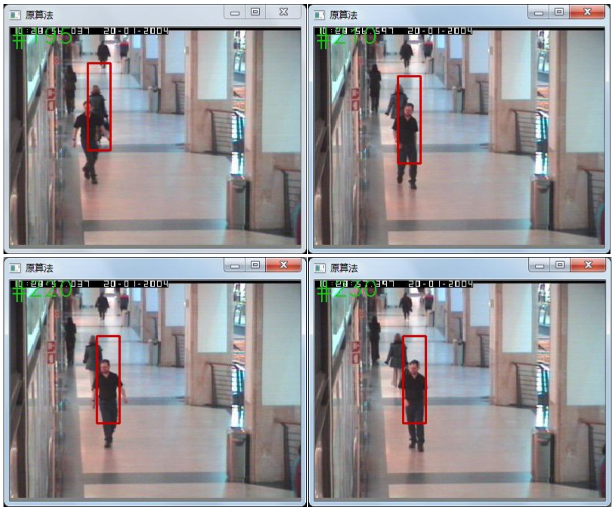


Figure 4. Walking2 sequence of the original algorithm 195,210,220,230 frames

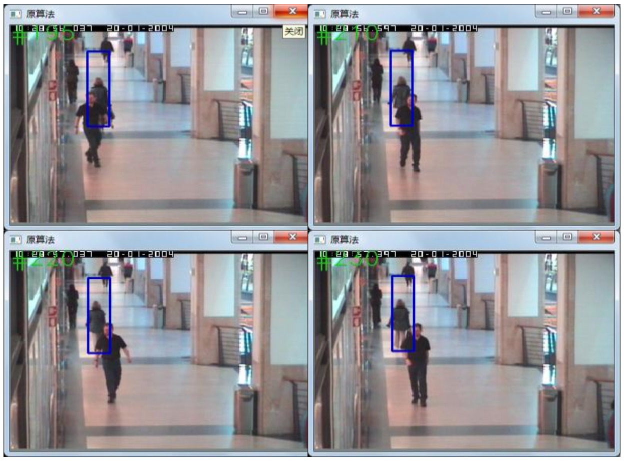


Figure 5. Walking2 sequence of the improve algorithm 195,210,220,230 frames

B. *Date Analysis*

After the above comparison using the sequence image, the following two kinds of algorithms for quantitative analysis, using two analysis methods. The first is the algorithm running speed comparison table 1, the unit is the number of transmission frames per second fps. Similar to the theoretical analysis, the speed of this algorithm is slightly lower than the original algorithm. The improvement scheme proposed in this paper basically maintains the real-time performance of the original algorithm. As can be seen from the data, for different image sequences are reduced, but the reduction is not significant.

The second is to track the success rate, the public test set to provide the tracking target real location and algorithm tracking position to compare. In the same sequence of test environment, tracking success rate has greatly improved, and enhanced the robustness of the algorithm.

TABLE I. ALGORITHM RUNNING SPEED COMPARISON

Sequence	**CT**	**Improved**
Box	69fps	64fps
Walking2	70fps	66fps

TABLE II. TRACKING SUCCESS RATE COMPARISON

Sequence	**CT**	**Improved**
Box	21.53%	73.55%
Walking2	20.15%	32.31%

V. CONCLUSION

The algorithm is mainly based on the compression-aware tracking algorithm, which is based on the improvement of the algorithm. In order to maintain the excellent real - time performance, the Kalman filter is introduced to improve the tracking effect of the tracking target scale transformation and partial occlusion by simple scale transformation. At the same time, it improves the parameter updating mechanism of the algorithm classifier, effectively suppresses the iterative accumulation of errors and improves the robustness of the algorithm. Experiments show that this algorithm is effective.

REFERENCES

[1] Donoho D. Compressed sensing [J]. IEEE Transactions on Infor-mation Theory, 2006, 52(4): 1289-1306.

[2] Candes E, Tao T. Near optimal signal recovery from random projections and universal encoding strategies [J]. IEEE Transactions on Information Theory(S0018-9448), 2006, 52(4) : 5406-5425.

[3] Achlioptas D. Database-friendly random projections : Johnson-Lindenstrauss with binary coins [J]. Journal of Computer and System Sciences(S0022-0000), 2003, 66 : 671‐687.

[4] Baraniuk R, Davenport M, De Vore R, et al. Wakin M. A simple proof of the restricted isometry property for random matrices [J]. Constructive Approximation(S0176-4276), 2008, 28 : 253‐263.

[5] Diaconis P, Freedman D. Asymptotics of graphical projection pursuit [J]. The Annals of Statistics(S0090-5364), 1984, 12(3) : 228‐235.

[6] Jain A, Nandakumar K, ROSS A. Score normalization in multimo-dal biometric systems[J]. Pattern Recognition, 2005, 38 (12):2270-2285.

[7] MAZINAN A H, AMIR‐LATIFI A. Applying mean shift, mo-tion information and Kalman filtering approaches to object tracking [J]. ISA transactions, 2012, 51(3) : 485‐497.

[8] ZHANG Kaihua, ZHANG Lei, YANG Ming-Hsuan. Real-Time Compressive Tracking [C]. Proceedings of the 12th European conference on Computer Vision, Florence, Italy, Oct 8-11, 2012, 3 : 866‐879.

Inferring Genome-Wide Gene Regulatory Networks with GPU or CPU Parallel Algorithm

Ming Zheng
Guangxi Colleges and Universities Key Laboratory of Professional Software Technology
Wuzhou University
Wuzhou, China
E-mail: 370505375@qq.com

Shugong Zhang
College of Mathematics
Jilin University
Changchun, China
E-mail: zhangsg@jlu.edu.cn

Mugui Zhuo
Guangxi Colleges and Universities Key Laboratory of Professional Software Technology
Wuzhou University
Wuzhou, China
E-mail: 756456050@qq.com

Guixia Liu*
College of computer science and technology
Jilin University
E-mail: liugx@jlu.edu.cn
*The corresponding author

***Abstract*—Expression of gene block, with the GPU parallel thread structure characteristic calculation, according to the structural characteristics of GPU thread design of double parallel mode, and the use of texture cache memory to achieve high efficiency; on the basis of CPU two level cache capacity of basic blocks further subdivided into sub blocks to improve the cache hit rate, the technology to reduce the number of memory accesses the use of data, reduce the thread migration in the core between the use of thread binding technology; according to the calculated capacity allocation of multi-core CPU and GPU CPU and GPU gene in the mutual information calculation task to calculate the load balance of CPU and GPU; in the design of the new threshold calculation algorithm based on the design and implementation of memory efficient construction of global gene control network CPU /GPU parallel algorithm. The experimental results show that compared with the existing algorithms, this algorithm speed is more obvious, and is able to build more large-scale global gene regulation Control network.**

Keywords-Genome-wide; Gene regulatory network; CPU /GPU cooperative computing; Efficient access cache; Parallel algorithm

I. INTRODUCTION

With the complete genome sequence of the human genome work sketch, multiple model organisms, after genomics genome era main focus from sequencing steering function research[1]. Analysis of gene expression microarray technology makes the establishment of global gene regulatory networks become possible, but the construction of gene regulatory network is very difficult[2]: every eukaryotic organisms have tens of thousands of genes, leading to the gene regulatory network to build a special complex; there is no model of a mature method, from gene expression analysis of gene regulatory relations spectrum map; there are a lot of noise and affect the gene expression significantly, increased the difficulty of constructing gene regulatory networks.

At present the construction of gene regulatory network model are: Bayesian network model[3] and mutual information model[4]. The Bayesian network model into directed acyclic graph model and hidden Markov chain to describe the relationship between Bayesian network variables and interactions, to construct regulatory network models. However, the Bayesian model of exponential time complexity in the construction of large-scale global the efficiency of gene regulatory network is very low[5]. Butte and IKohane proposed the use of mutual information as the detection of gene regulation relationships between complex tools, experiments show that the network model based on mutual information in the construction of regulatory network quality and time complexity and has obvious advantages[6]. The mutual information algorithm for constructing gene regulatory networks based on mostly serial algorithm the construction control network of approximately one thousand genes. These serial algorithms can only eukaryotic Creatures generally consist of tens of thousands of genes. The establishment of global gene regulatory network requires 109 orders of number of mutual information calculation.

This proposed algorithm is a global gene regulatory network platform design and implementation of multi-core CPU/GPU[7] in the parallel collaborative heterogeneous computing, the main contributions are as follows: the parallel construction of gene regulatory network model design, the design and implementation of the new regulatory threshold selection algorithm; design and implementation of the CPU and GPU memory efficient parallel computing gene the mutual information algorithm.

II. Construction of Gene Regulatory Network Model Based on Mutual Information

A. Mutual information estimation

Mutual information measures the correlation between two event sets, and the mutual information of the two events X and Y is defined as[8]:

$$I(X,Y)=H(X)+H(Y)-H(X,Y) \qquad (1)$$

H(X,Y) is the joint entropy in Eq. (1). The H(X,Y) can be shown as below:

$$H(X,Y)=-\sum_{x\in X,y\in Y}P(x,y)\log P(x,y) \qquad (2)$$

Where P(x, y) is the joint probability of X and Y, mutual information can be expressed as:

$$I(X,Y)=\sum_{x\in X,y\in Y}P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \qquad (3)$$

Mutual information I (X, Y) is a function of probability, can be estimated by using mutual information kernel function[9]. X n samples of known variable value, the variable X probability density function f (x) kernel function estimation:

$$\hat{I}(x)=\frac{1}{nh}\sum_{i=1}^{n}K(\frac{x-x_i}{h}) \qquad (4)$$

K is called the kernel function, h for the window width or smooth parameters. The window width parameter is usually equal to:

$$h\approx[\frac{4}{(d+2)}]^{\frac{1}{d+4}}zn^{-\frac{1}{d+4}} \qquad (5)$$

Where d is the dimension of the data set, z is the standard deviation of the sample data. By Eq. (3), I (X, Y) of the estimated value is shown as below:

$$\hat{I}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}\log\frac{\hat{f}(x,y)}{\hat{f}(x_i)\hat{f}(y_i)} \qquad (6)$$

The KSTest of the gene expression data in the gene expression profile shows that the kernel function is chosen to be normal distribution, so the kernel function is selected by Gauss function:

$$K(x)=(\frac{1}{\sqrt{2\pi}})^{e^{-\frac{x^2}{2}}} \qquad (7)$$

Finally, the estimated I(X,Y) can be obtained as below:

$$\hat{I}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}\log\frac{n\sum_{j=1}^{n}e^{-\frac{n^{2/6}}{2}[\frac{(x_j-x_i)2}{z_1^2}-\frac{(x_j-x_i)(y_j-y_i)}{z_1z_2}+\frac{(y_j-y_i)2}{z_2^2}]}}{\sum_{j=1}^{n}e^{\frac{-(x_j-x_i)2}{2(\frac{4}{3n})^{\frac{2}{5}}z_1^2}}\sum_{j=1}^{n}e^{\frac{-(y_j-y_i)2}{2(\frac{4}{3n})^{\frac{2}{5}}z_2^2}}} \qquad (8)$$

B. Mutual information estimation parallel model

The expression is subdivided into basic block gene; secondly, the computing tasks allocated to CPU and GPU cooperative computing; thirdly, the basic block is further subdivided into sub blocks, the multi-core parallel computing effective caching; finally, the design of GPU terminal two layers diagonal parallel computation, to achieve efficient access.

First of all the basic blocks, each GPU parallel diagonal matrix calculation results on all the matrix blocks, the block matrix is the two gene expression profile of basic block calculation, each block corresponding to the two basic block calculation. Then to base because the unit, each GPU parallel computing all the genes in a diagonal line inside the matrix blocks on the value of mutual information calculation, each thread corresponds to a pair of genes. In order to make GPU a parallel matrix block diagonal can be calculated for all genes on the mutual information value, set up the basic block containing gene number is equal to the number of threads in a thread block. The GPU parallel computing strategy agreement CUDA thread structure the characteristics, can increase the utilization rate of hundreds of core processing in GPU.

Each diagonal parallel computing all the genes on the value of mutual information, every expression of the need to calculate the matrix blocks gene basic blocks are different, such as the (i, j), the calculation of matrix blocks to gene expression of No. I and No. J basic block basic block spectrum expression. Two copies of this gene for data storage, you can make each diagonal parallel computing have no access conflict.

C. Determination of the regulation relationship

The threshold is an important parameter to control whether the evaluation of two gene regulation relationship, accurate determination of this parameter is difficult. This paper calculates the mutual information between the 200 gene values, and these values are sorted before 1000 increments (two mutual information and each adjacent difference value plotted) as shown in the figure.

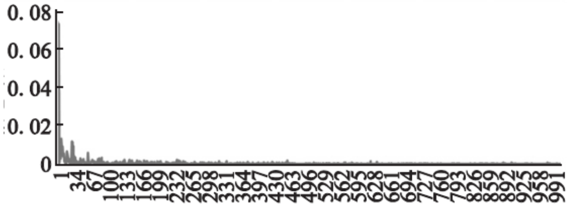


Figure 1. Incremental curve for mutual information

From the picture we can see that the change of mutual information was large and then leveled off, which shows obvious inflection point mutual information curve. At the same time, Fig .1. Curve jitter phenomenon is obvious, and the first change of mutual information is too large. So how to eliminate the chattering phenomenon to accurately identify the inflection point (threshold) is one of the key problems. This paper competition scoring system, using the following method to eliminate the effects of jitter to accurately find the threshold value: the sort of mutual information of all genes calculated; the minimum mutual information value, calculate the increment between them, get rid of one of the largest and the smallest one for the rest of the average increment, increment value; the increment threshold for alpha times to average, of which 0.001 < alpha < 0.1; if the increment between 10 consecutive mutual information are small In the incremental threshold, the corresponding mutual information value is the desired control threshold.

But the mutual information on all eukaryotic gene values about hundreds of millions, of mutual information values for all the sort of large computational complexity; and the position corresponding to the inflection point threshold should be in the mutual information value that is relatively small. Therefore, this paper calculates the threshold to remove the mutual information minimum mutual information a value of 5%, and then refer to the "two search" threshold selection method from the 5% in the value of mutual information.

III. ANALYSIS OF THE PROPOSED PARALLEL ALGORITHM

The main idea of the algorithm: gene expression profiling is subdivided into basic blocks, with basic blocks on the diagonal parallel computing; distributed computing times for CPU and GPU; according to the two level cache multi-core structure of the capacity of basic blocks further subdivided into sub blocks, and the next time to calculate the required data to prefetch cache GPU; end take double diagonal parallel computing, the use of texture memory bound data.

Global gene regulatory network result matrix is very large, the construction of gene regulation network of the 50 thousand genes the matrix size is about 10GB, a single GPU memory to store the entire result matrix, so take part the result of each GPU storage, then summary results.

Algorithm 1. Constructs a parallel algorithm of CPU and GPU for global gene regulation network

```
Gene expression profile
Gene regulatory network
Begin
(1) read the gene expression profile, and according to the GPU thread size BlockRowCount, calculation of NumBlock and calculation of basic blocks round ComputeCount;
(2) calculation is assigned to CPU and each GPU round ComputeCountCPU calculation
And ComputeCountGPU;
(3) do steps (3.1), (3.2) in parallel
(3.1) call CPU parallel computing mutual information algorithm (algorithm 2);
(3.2) call multi GPU parallel computing mutual information algorithm (algorithm 3);
(4) summary of the result matrix returned by multiple GPU;
(5) the threshold value calculation algorithm (algorithm 4) is used to calculate the threshold value and the threshold is used to filter the mutual information matrix;
(6) the mutual information matrix of gene was analyzed by DPI, and the control network was further simplified;
End
```

The algorithm 1 is mainly based on GPU thread structure partition, then according to the calculation ability of CPU and GPU will calculate the corresponding rounds assigned to CPU and GPU, in order to achieve load balance.

Algorithm 2.CPU parallel computing gene pair mutual information algorithm

```
Input: gene expression profile, ComputeCountCPU, CPU thread number thread-NumCPU
Output: CPU end mutual information calculation result matrix
Begin
(1) calculate the number of lines of the w block, the basic block is further divided into sub blocks, the number of sub blocks are SubNum, and calculate the parameters of k = w /threadNumCPU;
(2) with the instruction prefetch expression basic block prefetch to level three cache memory gene number zeroth, and a copy of this, two pieces of data were recorded as basic blocks A and B;
(3) for I = 0 to do BlockNum1
(3.1) for J = 0 to do ComputeCountCPU1
Do steps (3 1.1) ~ (~ 3) in parallel
(3.1.1) with the three level cache prefetch instruction No. I No. zeroth block gene expression profile in a basic block read to the two level cache, and a copy of this, two blocks are respectively denoted as sub block SA and sb;
(3.1.2) for Si = 0 to do SubNum1
(3.1.2.1) = 0 for SJ to do SubNum1
Do steps (3 1 1.1) ~ (3 1 2 1.2) in parallel ()
(3.1.2.1.1) = 0 to for TID par-do TheadNumCPU1
For SWI = 0 to do W1
For swj = tid* to (TID + 1) * k do K
Begin
The Y (SWI + swj)% w gene Y of the swj gene X and the sub block sb in the sub block SA is read into the primary cache from the two level cache; the mutual information of the X and the;
End
(3.1.2.1.2) if (SJ + 1) < SubNum based prefetch instruction reads the gene number I from the three level cache spectrum basic block in the SJ + 1 chant block to the two level cache replacement block sb expression;
End for
(3.1.2.2) if (Si + 1) < SubNum based prefetch instruction reads the gene number I from the three level cache spectrum basic block in the Si + 1 chant block to the two level cache replacement block SA expression;
End for
(3.1.3) with a prefetch instruction from main memory into the (I + j + 1) expression of basic blocks to level three cache replacement basic block B%BlockNum gene;
End for
(3.2) with prefetch instructions read from main memory I + expression of basic block to level three cache replacement basic block A 1 gene;
End for
End
```

Algorithm 2 according to the cache capacity of CPU, made a further subdivision of the basic block into several sub blocks, and then prefetch the basic block, sub block to level three, level two cache, the number of accessing main memory was significantly reduced. The partition can make the three level cache can be transferred to the 4 basic block, divided the sub block can make the two level cache can be transferred to 4 sub blocks, so the three level cache and level two cache can accommodate the next calculation calculation and the data needed to achieve zero loss. At the same time the use of "cache latency hiding" model, computing and memory access overlap, forming multilevel pipeline model, make the calculation the process has been accelerated.

Algorithm 3 multi GPU parallel computing gene pair mutual information algorithm

```
Input: gene expression profile, ComputeCountGPU, GPU thread block size
Output: the mutual information result matrix for each GPU gene
Begin
For each GPU do in parallel
(1) specify a calculation GPU;
(2) from the memory transfer of gene expression data to GPU, and the use of 2D texture structure of these data is bound to the texture memory;
(3) for I to do in parallel BlockNum1 = 0
For J = 0 to do ComputeCountGPU1
For Ti = 0 to do in parallel BlockRowCount1
For TJ = 0 to do BlockRowCount1
Begin
The number of Ti (Ti + J - 1) gene expression X (I + Tj)% BlockRowCount gene in the basic block of the gene expression profile of I gene was studied. The mutual information between X and Y was calculated by Y;
End
(4) from the GPU memory to memory transfer matrix results;
End for
End
```

IV. Experiment

A. Experimental and Experimental Data

The experimental data from the public gene expression database GEO[10], this paper used two groups of gene expression data: contains 32996 genes and each gene has 25 sample data set GSE7431, contains 54675 genes and each gene has 143 sample data sets GSE22148.

The experimental platform for the 2 XEON E5620 2 4GHz 4 core Intel processor and 4 GPU (4 × Nvidia Tesla C2050 3GB) of the multi-core computer, the memory capacity of 12GB, sharing the three level cache capacity of 12MB, the two level cache capacity of each core private 512KB, a cache capacity of 64KB, operation the system is running red Hat Enterprise 5 Linux, OpenMP and CUDA using C language programming.

B. Experimental Results and Analysis

For data set GSE7431, run 1, 2, 3 and 4 GPU, respectively, each thread block in the operation of the thread, Fig .2 gives the algorithm in this paper, the parallel computation of the 3 genes on the time required for mutual information:

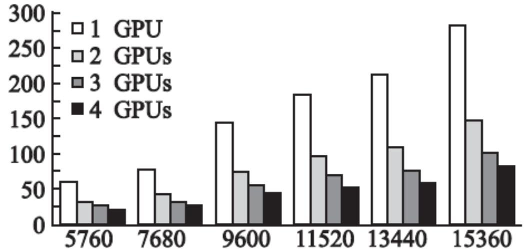


Figure 2. Required time to execute Algorithm 3 running GPUs with different number to compute mutual information

The experimental results show that the more GPU algorithm operation, less calculation is needed for the gene mutual information time; in addition, can also see that the computation time is about running n GPU. single run this shows that GPU can effectively enhance the performance of parallel computing, algorithm, this algorithm is suitable for the operation of 3 in GPU system, with good scalability.

Table 1 gives the data set GSE7431 gene on serial computing mutual information algorithm, this algorithm 2, algorithm 3 run 64 threads running 4 GPU and each thread block has 192 threads, respectively calculate the genes required to mutual information time, speedup and parallel algorithm 2 and 3 obtained.

TABLE I. REQUIRED TIME TO COMPUTE MUTUAL INFORMATION USING SERIAL ALGORITHM, ALGORITHM 2 AND ALGORITHM 3

number	time	Algorithm 2		Algorithm 3	
		time	Speedup	time	Speedup
960	291	257	6	4698	23
2240	1587	1	32	8673	48
3520	3924	81	73	7423	53
4800	7311	93	130	704	55
6080	11724	8	216	521	54
7360	17203	8	308	268	55

From Table 1, the experimental results show that the parallel multi core CPU and GPU parallel computing of genes has accelerated effect on mutual information. For the 960 gene data for smaller, because the parallel overhead of CPU parallel algorithm running time accounted for a larger proportion, so the acceleration effect is not obvious, the speedup is only 23; when the data size increases to a certain extent, multi-core CPU parallel algorithm of acceleration is relatively stable, about 55. 960 genes for small data size, GPU parallel algorithm and computation time than CPU in parallel, it is because the GPU communicates with the CPU time of a larger proportion, influence the performance of the algorithm when; the data size increases to a certain extent, the speedup increases rapidly, the GPU parallel algorithm's advantage is obvious, this shows that the GPU parallel algorithm is suitable for large-scale data gene The calculation of mutual information.

For the GSE7431 data set, each thread block has 192 threads, the next page is shown in Fig and CPU algorithm 3 run 64 threads and GPU threads block has 192 threads parallel time calculation algorithm in this paper 1 gene on the mutual information.

V. CONLUSION

CPU and GPU proposed the parallel computing of the gene mutual information algorithm to build more large-scale global gene regulation networks and significantly shorten the construction of global gene regulatory network is the most complex gene computation time of mutual information, because it is on the gene expression profile of block, according to the structural characteristics of GPU parallel thread computing according to the structural characteristics of GPU, design the double thread parallel mode, and the use of texture cache memory to achieve high efficiency; based on nuclear CPU cache, the basic block further subdivided into sub blocks to ensure cache zero loss, take the technology to reduce the number of memory accesses the data pre, reduce the thread migration in the core between the use of thread binding technology; the task to achieve CPU and GPU load balancing through the rational allocation of the CPU and GPU calculation. The next step will be the reference of community discovery thoughts on global gene Module partition method of control network

ACKNOWLEDGMENT

This work was supported by grants from The National Natural Science Foundation of Chi-na (No. 61502343, No. 61373051, and No. 61402423), China Postdoctoral Science Foun-dation funded(No. 2016M590260), the Guangxi Natural Science Foundation (No. 2015GXNSFBA139262), the Science Research Funds for the Guangxi Universities (No. KY2015ZD122), Guangxi Colleges and Universities Key Laboratory of Professional Software Technology, Wuzhou University.

REFERENCES

[1] Carter, M.Q.: ‘Decoding the Ecological Function of Accessory Genome’, Trends Microbiol., 2017, 25, (1), pp. 6-8

[2] Fujii, C., Kuwahara, H., Yu, G., Guo, L.L., and Gao, X.: ‘Learning gene regulatory networks from gene expression data using weighted consensus’, Neurocomputing, 2017, 220, pp. 23-33

[3] Fan, Y., Wang, X., and Peng, Q.K.: ‘Inference of Gene Regulatory Networks Using Bayesian Nonparametric Regression and Topology Information’, Computational and Mathematical Methods in Medicine, 2017

[4] Chen, C., and Yan, X.F.: ‘Optimization of a Multilayer Neural Network by Using Minimal Redundancy Maximal Relevance-Partial Mutual Information Clustering With Least Square Regression’, IEEE Trans. Neural Netw. Learn. Syst., 2015, 26, (6), pp. 1177-1187

[5] Thorne, T.: ‘NetDiff - Bayesian model selection for differential gene regulatory network inference’, Scientific Reports, 2016, 6

[6] Kurt, Z., Aydin, N., and Altay, G.: ‘Comprehensive review of association estimators for the inference of gene networks’, Turkish Journal of Electrical Engineering and Computer Sciences, 2016, 24, (3), pp. 695-U1401

[7] Wei, R., and Murray, A.T.: ‘A parallel algorithm for coverage optimization on multi-core architectures’, Int. J. Geogr. Inf. Sci., 2016, 30, (3), pp. 432-450

[8] Zu-yun, F.: ‘Information theory: basic theory and application’ (2007, 2nd Edition edn. 2007)

[9] Yan, X.Y., Zhang, S.W., and Zhang, S.Y.: ‘Prediction of drug-target interaction by label propagation with mutual interaction information derived from heterogeneous network’, Molecular Biosystems, 2016, 12, (2), pp. 520-531

[10] Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N.G., Robertson, C.L., Serova, N., Davis, S., and Soboleva, A.: ‘NCBI GEO: archive for functional genomics data sets-update’, Nucleic Acids Res., 2013, 41, (D1), pp. D991-D995

Association Rule Mining Based on Estimation of Distribution Algorithm for Blood Indices

Xinyu Zhang
College of Information Science and Engineering
Northeastern University
Shenyang, China
E-mail: zhangxinyu1995@126.com

Guanghu Sui
College of Information Science and Engineering
Northeastern University
Shenyang, China
E-mail: suiguanghu@foxmail.com

Botu Xue
College of Information Science and Engineering
Northeastern University
Shenyang, China
E-mail: xuebotu1994@163.com

Jianjiang Cui*
Inst. Intelligent Systems
Northeastern University, Shenyang, China
E-mail: cuijianjiang@ise.neu.edu.cn
*Corresponding Author

***Abstract*—To come over the limitations of Apriori algorithm and association rule mining algorithm based on Genetic Algorithm (GA), this paper proposed a new association rule mining algorithm based on the population-based incremental algorithm (PBIL), which is a kind of distribution estimation algorithms. The proposed association rule-mining algorithm keeps the advantages of GA mining association rules in coding and the fitness function. Through using probability vector possessing learning properties to update the population, the algorithm increases the convergence speed and enhances the searching ability, compared to GA. In the experiment of mining association, rules in blood indices data, PBIL algorithm performs better not only in running time, convergence speed, but also achieve better searching results. Meanwhile, this paper proposed a parallel algorithm for association rule mining based on PBIL and designed a system architecture based on cloud computing for blood indices analysis, providing a good example to apply the new algorithm to cloud computing.**

Keywords-Distribution estimation algorithm; Probability vector; Blood indices; Parallel algorithm; Cloud computing

I. INTRODUCTION

Association rule mining is an important branch of data mining. By collecting many records of items in the database then analyzing them, the valuable relationships between huge amounts of data can be found [1]. The significance of the association rules analysis is greater in medical data than in other areas. By mining the medical data, the potential relationships between various diseases and various health indicators can be found, serving for medical research and disease diagnosis [2]. Apriori algorithm is the most typical algorithm for association rule mining. Traditional Apriori algorithm needs to scan the database for many times to generate vast candidate sets, leading to poor extensibility of Apriori algorithm. To overcome the weakness above, some scholars put forward a theory using intelligent optimization algorithm to mine the association rules. In 2004, Li Ying [3] put forward the application of generalized genetic algorithm in Apriori algorithm improving. At first, it uses Apriori to search partial association rules, then it uses genetic algorithm to search global association rules. In this way, the times of traversing the database can be reduced. In 2012, Shiwei Chen [4] put forward a method of association rule mining based on interest measure and genetic algorithm, improving the quality of association rule mining. In 2016, Donghao Xu [5] put forward a method of association rule mining based on improved particle swarm optimization algorithm, verifying the advantage of particle swarm optimization on association rule mining compared with genetic algorithm.

With the rapid development of computer technology, cloud computing has become a direction for the future development of distributed computing. MapReduce programming frame put forward by Google is a representative technology of cloud computing. It is suitable for distributed processing of large-scale datasets and has very high computational efficiency [6]. Therefore, some scholars put forward a method of association rule mining based on Hadoop and other cloud computing technology. They also put forward some parallel algorithms for association rule mining. In 2011, Zhang Sheng [7] put forward an Apriori algorithm based on cloud computing. It deploys MRM-Apriori algorithm in MapReduce frame and has good effect in speed.

Based on the research, an association rule mining algorithm based on distribution estimation (PBIL) is proposed in this paper. Compared with Apriori and GA association rule mining algorithm in experiment, the new algorithm is proved to be effective. Meanwhile, an parallel algorithm based on this algorithm is designed, which is suitable for MapReduce frame. In addition, a realization plan for blood indices analysis system based on cloud computing is introduced in this article.

II. The Classical Association Rule Mining Algorithm

Apriori algorithm is one of the most typical algorithms in the data mining field. It uses a method called iteration of layer by layer to produce high dimension frequent item sets from low dimension frequent item sets. Then the association rules can be produced from frequent item sets [8]. The specific mathematical model is in literature 8.

Support degree in Apriori algorithm is defined as follows: the number of transactions of the entire transaction set is m, and there are n transactions containing the item set, then the support degree of the item set is n/m. If the item set A exists, the support degree of the item set is supp(A).

The candidate set in Apriori algorithm is generated layer by layer, and only after fully scanning the database will the frequent item set of this layer be produced. Therefore, if the database is very large, this work will cost a lot of memory resources, reducing the efficiency of the algorithm.

III. Association Rule Mining Based on Genetic Algorithm (GA)

A. Association Rules Model Based on GA

Based on the analysis of the Apriori algorithm, association rule mining is mainly divided into two parts. First, find frequent item sets in the transaction database. The second is to generate association rules based on the frequent item sets which are found [9]. And the workload of the former is the larger one, which is the direct cause of low efficiency of Apriori. Therefore, genetic algorithm (GA) can be used to realize the global search for frequent item sets.

Genetic algorithm is an effective global optimization algorithm. With its binary coding mode, using genetic operators to evolve population, it is able to keep extracting frequent item sets from the transaction database at a rapid speed, avoiding the operations like join and prune which need to frequently scan the database, improving the computing speed and mining precision. The concrete implementation steps are shown in fig. 2.

B. The Weakness of Association Rule Mining Based on GA

Genetic algorithm will save the individuals, which meet the requirements of frequent item sets in each generation to the next generation when it is mining association rules. And the individuals which meet the requirements will be removed from the previous generation. The main purposes of this population selection method are to reserve the excellent individuals and to keep the population diversity. But these two requirements are contradictory. If the excellent individuals reserved are excessive in each generation, the population diversity will decrease, prone to lead to a prematurity phenomenon. As a result, in association rule mining, the search of frequent item sets will be incomplete, and the extracted association rules will be partial. If the excellent individuals reserved are not enough in each generation, the convergence speed of the algorithm will be reduced and the computing time will be longer. That will affect the advantage of association rule mining based on genetic algorithm in speed.

IV. Association Rule Mining Algorithm based on Estimation of Distribution Algorithm (EDA)

A. Description of EDA

To come over the disadvantages of genetic algorithm mining association rules, EDA can be used. EDA is a kind of evolutionary algorithms developed by the genetic algorithm. It first selects samples from the optimal population and extracts information. Then it uses the information to build proper probability module. At last, it updates the population to increment individuals with more fitness until the end condition. In the way it can maximum the individuals' quantity and keep the population diversity [10]. At the same time, EDA can select new solutions by probability distribution to obtain the optimal solutions with less iteration times. It can effectively prevent the local optimization and precocity in GA when dealing with higher order or long-distance tectonic block problems [11].

B. PBIL Probability Module

When handling the problem, which owns mutual independent variables, PBIL algorithm, a typical form of EDA, can be used. PBIL algorithm is mainly applied to binary-code optimization problem. PBIL collects the data recording the values of variables, whose value is ruled as 0 or 1, to build the probability vector. Then it uses the probability vector to estimate the one-dimensional edge distribution. Assuming that a binary gene population with N gene positions (mutual independent variables) and M individuals is existing and the population can evolve continuously, the gene population on the tth generation can be expressed as:

$$g_i^{X_t^j} = \begin{cases} 0 \\ 1 \quad (i=1,...,M,\ j=1,...,N) \end{cases} \tag{1}$$

Where t represents the evolutional generation number of the gene population, $X_t^j\left(j=1,...,N\right)$ represents the jth gene position (independent variable) at the tth generation, $g_i^{X_t^j}$ represents the code of X_t^j th gene position (variable) of the ith individual at the tth generation.

Then we use the code condition of every variable at the tth generation to generate the probability vector:

$$P_t = \left(P_t\left(X_t^1\right) \quad P_t\left(X_t^2\right) \quad P_t\left(X_t^3\right) \quad ,..., \quad P_t\left(X_t^N\right)\right) \tag{2}$$

Then we count the amount of the individuals whose designated gene position (variable) are of code 1 and calculate the percent in total M individuals at current gene population. The percent obtained is equal to the distribution probability of the designated variable as follows:

$$P_t\left(X_t^j\right) = \frac{\sum_{i=1}^{M} g_i^{X_t^j}}{M}\left(g_i^{X_t^j} == 1\right) \tag{3}$$

When using the EDA, it first generates a random original population and figures out the fitness value of every individuals of the population. Then it ranks all individuals in order of corresponding fitness value. Individuals with greater fitness values will be seen more advanced. Then truncation is adopted to select advanced individuals of certain amount. The rate of truncation is named as selerate. So the amount of advanced individuals m can be expressed as:

$$m = selerate \cdot M \qquad (4)$$

Therefore, the advanced population is made up of the first ranked m individuals of the original population.

Through equation (3) PBIL algorithm gains the probability vector by those advanced individuals. Then it takes samples basing on the probability vector and the next-generation population is obtained. At the same time to make probability vector describe the probability distribution of the advanced individuals with faster speed and more accurate quality, PBIL algorithm adopts the Heb rules from machine learning theory to update the probability vector, which means that the probability distribution of each variable is adjusted linearly at a certain learning speed[12] as equation (5) shows:

$$P_{t+1}\left(X_t^j\right) = (1-\alpha)P_t\left(X_t^j\right) + \alpha\frac{1}{m}\sum_{i=1}^{m} g_i^{X_t^j} \qquad (5)$$

Where m represents the amount of advanced individuals selected at the tth generation.

The process of sampling from the probability vector can be described as generating a random number ranging from 0 to 1. If the number is greater than the probability vector corresponding to a certain individual gene position, which means, the binary value of the gene position is 1, otherwise 0.

C. Codes of Individuals and Item Set

When dealing with the problem of mining association rules, EDA adopts binary code also. That is, when a patient owns an abnormal blood index, the value of the index is set as 1. If the index is normal, the value is set as 0 instead.

D. Selection of Fitness Function

Fitness function is designed to reflect the frequency of the item set and recognize the frequent item set. Since the criteria of being frequent item set is that the support level of the item set is greater than the minimum support level, the fitness function can be defined as:

$$fitness\left(g_i^{X_t^j}\right) = \frac{Supp\left(g_i^{X_t^j}\right)}{MinSupp} \qquad (6)$$

Where $g_i^{X_t^j}$ represents the ith individual at the ith generation. The name of the fitness function is fitness. $Supp\left(g_i^{X_t^j}\right)$ represents the support level of the item set corresponding to a certain individual. $MinSupp$ represents the minimum support level which is given by user.

If an individual owns fitness value greater than 1, it illustrates the item set corresponding to the individual has its support level greater than the minimum support level. The item set is frequent item set. The individual will be reserved as a member of advanced population with updating the probability vector. If the fitness value is less than 1, it means that the item set corresponding to the individual is not frequent. Then the individual is eliminated directly.

E. Procedures of Mining Association Rules by PBIL

Based on the analysis above, specific steps of association rule mining based on PBIL algorithm is shown in Fig .1.

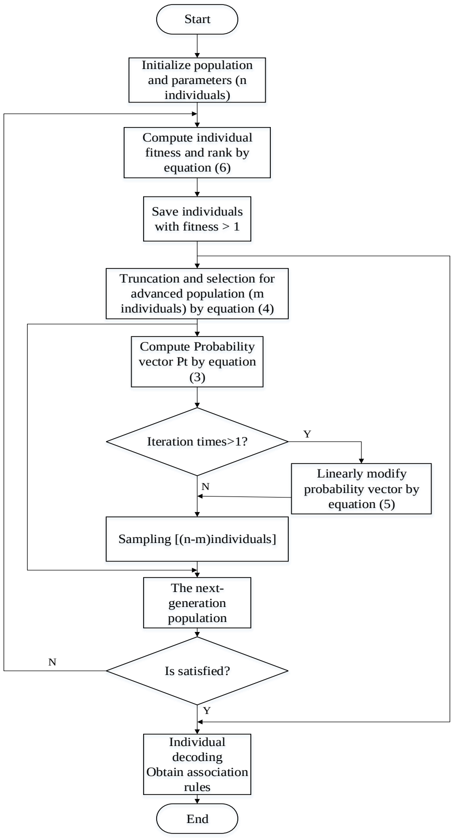

Start
Initialize population and parameters (n individuals)
Compute individual fitness and rank by equation (6)
Save individuals with fitness > 1
Truncation and selection for advanced population (m individuals) by equation (4)
Compute Probability vector Pt by equation (3)
Iteration times>1?
Y
Linearly modify probability vector by equation (5)
N
Sampling [(n-m)individuals]
The next-generation population
N
Is satisfied?
Y
Individual decoding Obtain association rules
End

Figure 1. Flow chart of PBIL mining association rules

V. Analysis of Experimental Results

A. Construction of the Transaction Database

All the data in this experiment are from anonymous blood routine laboratory sheets provided by a second grade hospital. Blood routine laboratory sheets provide 9 test indices [13], including white blood cell count (WBC), neutrophil (NE), lymphocyte (LY), monocyte (MO), eosinophil (EOS), basophil (BASO), red blood cell count (RBC), hemoglobin (Hb) and hematocrit (HCT). 255 laboratory sheets were randomly selected to build transaction database. The data in each laboratory sheet is regarded as one item set of transaction database. And each item of the item set is encoded referring to the encoding rules in 4.3 and the test result in the laboratory sheet: Normal index encoding is 0, and abnormal index encoding is 1.

B. The Experiments and Results Analysis

In order to verify the advantages of PBIL algorithm for mining association rules, classical Apriori algorithm and association rule mining algorithm based on genetic algorithm (GA) were compared with PBIL algorithm in the experiments. The three algorithms were compared with each other in three aspects: effect of mining association rules, the time of extracting frequent item sets and the convergence of PBIL.

Referring to the analysis of 3.1, association rule mining model based on GA can be designed. The specific steps are shown in Fig .2.

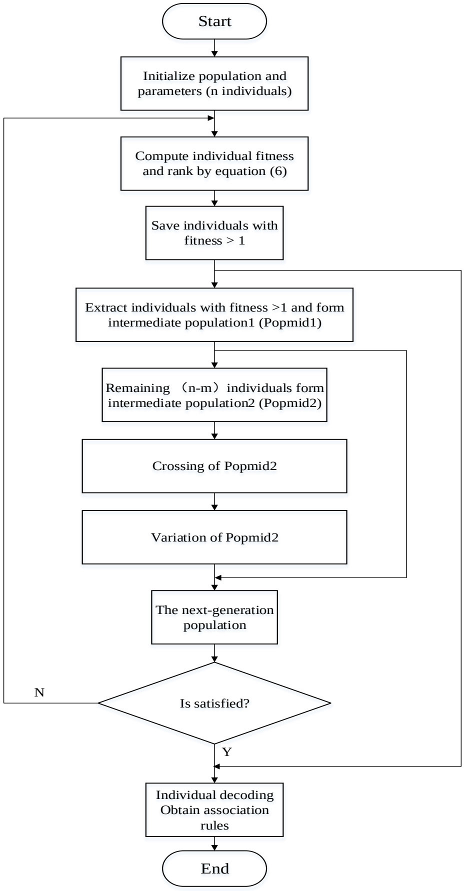

Start
Initialize population and parameters (n individuals)
Compute individual fitness and rank by equation (6)
Save individuals with fitness > 1
Extract individuals with fitness >1 and form intermediate population1 (Popmid1)
Remaining （n-m）individuals form intermediate population2 (Popmid2)
Crossing of Popmid2
Variation of Popmid2
The next-generation population
N
Is satisfied?
Y
Individual decoding Obtain association rules
End

Figure 2. Flow chart of GA mining association rules

Referring to the blood indices data in 5.1, based on Win10 system and Intel Core i5 processor, the algorithm above can be translated to programming in MATLAB 2015a.

C. Analysis of association rules

The transaction database (255 transactions, 9 data items) in 5.1 is calculated by using association rule mining algorithm based on PBIL. The set points are as follows: Minimum support is 0.12 (30/255); Minimum confidence is 0.7; Population size (Popsize) is 500; Iteration times (Iteration) are 100; Truncation selectivity (selrate) is 0.4; Learning rate (learnrate) is 0.1. Merging the similar rules of operation result, the final result is shown in Table I.

TABLE I. FOUND RULES

Rule number	***Rule premise***	***Rule result***	***Support level***	***Confidence level***
1	WBC,NE	LY	0.1294	0.8250
2	WBC,BASO	LY	0.1294	0.7174
3	WBC,Hb	HCT	0.1804	0.7302
4	NE,BASO	WBC	0.1216	0.8185
5	NE,RBC	Hb	0.1608	0.7885
6	NE,Hb	WBC	0.1412	0.8571
7	BASO,Hb	ESO	0.1216	0.7949
8	HCT,RBC	Hb	0.1294	0.8049

The result in Table I can be obtained in classical Apriori algorithm as well. Table I shows the incidence relations between each blood index. For instance, in Association Rule 1, patients with white blood cells, neutrophils and lymphocytes abnormal at the same time are the most common. Therefore, according to the association rule, patients with white blood cells abnormal can be told to prevent or treat diseases caused by abnormal lymphocytes, and vice versa.

D. Mining time comparison between algorithms

Based on the analysis in 3.1, association rule mining can be divided into two stages. The first stage is to find frequent item sets in the transaction database, which costs the main computing time. The second stage is to generate association rules based on the frequent item sets which are found. Parts of the three algorithms in the second stage are the same. Therefore, it is just enough to compare the three algorithms' time of searching for frequent item sets.

E. The relationship between the mining time and the number of transactions

In this experiment, the minimum support is 2/255, and the number of data items (indices) is 9. The three algorithms' computing time can be compared under the premise of searching for the same number of frequent item sets, by keeping changing the number of transactions. The setting parameters of PBIL and GA are as follows: Population size is 500; Iteration times are 100; PBIL truncation selectivity is 0.4; Learning rate is 0.1; GA crossover probability is 0.8; Mutation probability is 0.01. The change of the three algorithms' computing time is shown in Fig .3, and the number of frequent item sets which are found is shown in Table II.

In Fig .3, the changing number of transactions in the transaction set is used as abscissa. The computing time of algorithms is used as ordinate. The gray curve, orange curve and blue curve separately represent the trends of Apriori, GA and PBIL on computing time. In the condition of the same number of data items, the computing time of the three algorithms increases with increment of the number of transactions. The computing time of Apriori is the longest, surpassing PBIL and GA. The PBIL is based on the probability model to evolve population, and it has learning properties. Therefore, convergence of this algorithm is directional. GA is based on rules of crossover and mutation in nature to evolve population. So it has large randomness and doesn't have learning properties. As a result, PBIL has faster convergence speed than GA, which becomes more obvious with the increasing of data volume.

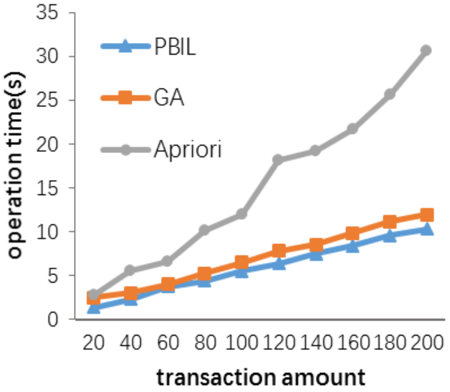


Figure 3. Operation time of algorithms at different transaction amount

TABLE II. FREQUENT ITEM SET'S AMOUNT FOUND IN DIFFERENT TRANSACTIONS' AMOUNT

Order	*Transactions' amount*	*Frequent item sets' amount*
1	20	95
2	40	197
3	60	205
4	80	250
5	100	253
6	120	331
7	140	332
8	160	336
9	180	356
10	200	373

F. *The relationship between the mining time and the number of data items*

In this experiment, the minimum support is 50/255, and the number of transactions is 255. The three algorithms' computing time can be compared under the premise of searching for the same number of frequent item sets, by keeping changing the number of data items (indices). The change of the three algorithms' computing time is shown in Fig .4, parameter setting and the number of frequent item sets which are found are shown in Table. III.

In Fig .4, in the condition of the same number of transactions, the computing time of the three algorithms increases with increment of the number of data items. The computing time of Apriori increases the most fast with the increment of data dimension. PBIL and GA obtain frequent item sets by searching for them, so the two algorithms are less influenced by data dimension. In addition, PBIL has much faster speed than GA. The reason is as follows: PBIL uses probability vector to evolve population, building a corresponding probability model for each variable of the individual. If an additional dimension is added to the data, a corresponding probability vector will be built. Each of the variables is mutually independent, evolving with its own probability vector, greatly reducing the affect caused by the increment of dimension. GA will strengthen the affect caused by the increment of dimension when additional dimensions are added to the data. It will constantly add high-dimensional data into the population to evolve because of its crossover and mutation. Therefore, PBIL algorithm is better than the former two algorithms.

TABLE III. FREQUENT ITEM SETS' AMOUNT FOUND IN DIFFERENT DATA SETS' AMOUNT

Order	*Data sets' amount*	*Frequent item sets' amount*	*Scale of population*	*Iteration times*
1	3	6	20	5
2	4	7	25	10
3	5	8	35	10
4	6	11	65	20
5	7	16	100	40
6	8	23	150	40
7	9	24	200	50

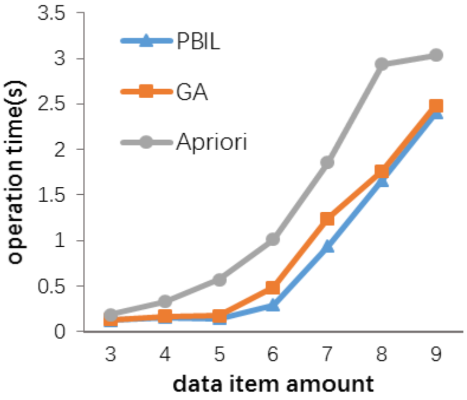


Figure 4. Operation time of algorithms at different data item amount

G. *Convergence of PBIL*

In this experiment, PBIL will be verified to have better convergence, compared with GA algorithm. The data with 9 transactions and 5 data items is used to experiment. The minimum support is 2/9. Population size is 15. Iteration times are 100. The other parameters are ditto. The best fitness value of each generation can be obtained by operating 20 times. The curve whose convergence speed is the fastest among the 20 operations of GA is used to compare with the convergence curve of PBIL. The result is shown in Fig .5. The maximum support of frequent item sets is 4.5. The two curves represent the two algorithms' ability of searching for frequent item sets with the maximum support. The figure shows that the best fitness value of each generation of PBIL algorithm reaches 4.5 first. GA is later than PBIL for at least 50 iteration periods. What's more, PBIL has found 8 frequent item sets, and GA has found 4. PBIL costs less time as well. Therefore, PBIL algorithm is better at convergence and searching ability.

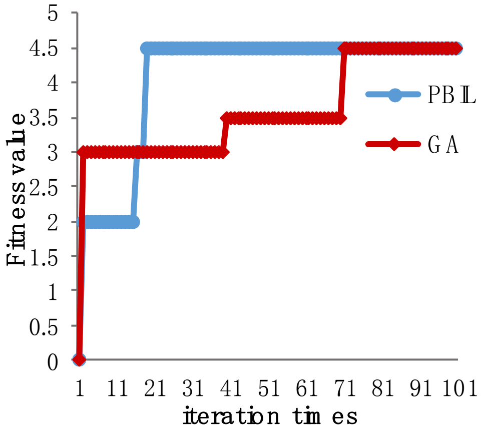


Figure 5. Astringency curve of PBIL vs. GA

VI. Mobile Analysis System of Blood Indices Based on Cloud Computing

A. *Parallel Algorithm of Association Rule Mining Based on PBIL*

The operation speed of the association rule mining algorithm based on PBIL depends on the population scale and iteration times. Therefore, we can consider decomposing the database, declining the scale of the transaction database and gather the results after parallel mining with PBIL algorithm. Basing on the thought above, the currently popular cloud-computing framework Hadoop [14] is used. Then we design the algorithm under the MapReduce framework inside Hadoop and use map function to execute data decomposition and mining. At last we use reduce function to gather the mining results. The framework of the algorithm is shown in Fig .6. Since each map function can achieve parallel computing, which means that searching all frequent item set costs approximately same time as searching a small database, the efficiency of the algorithm is greatly improved.

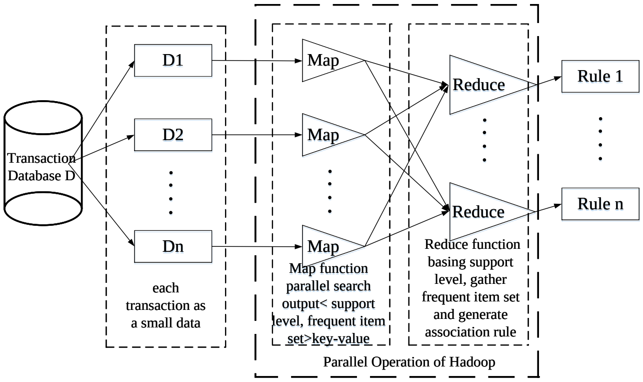

Transaction Database D
D1
D2
Dn
each transaction as a small data
Map
Map
Map
Map function parallel search output< support level, frequent item set>key-value
Reduce
Reduce
Reduce function basing support level, gather frequent item set and generate association rule
Rule 1
Rule n
Parallel Operation of Hadoop

Figure 6. Framework of parallel algorithm mining association rules

B. *Blood-Health Indices Analysis System Based on Cloud Computing*

The framework of the blood-health indices analysis system based on cloud computing is shown in Fig .7. Firstly, the system implements the algorithm in 6.2 by configuring the parallel-computing server cluster basing on Hadoop framework. The newly built database is connected with the hospital's database to update the data dynamically. To make it easier to use the system, an Android mobile application is developed to help analyze the indices of blood. Clients can upload abnormal indices to the server then the server will search for the indices, which form association rules with those indices uploaded according to the computing results. After that, it will read the referred illness symptoms and cure method. Eventually the server will send the information to the mobile clients to accomplish the online analysis of illness.

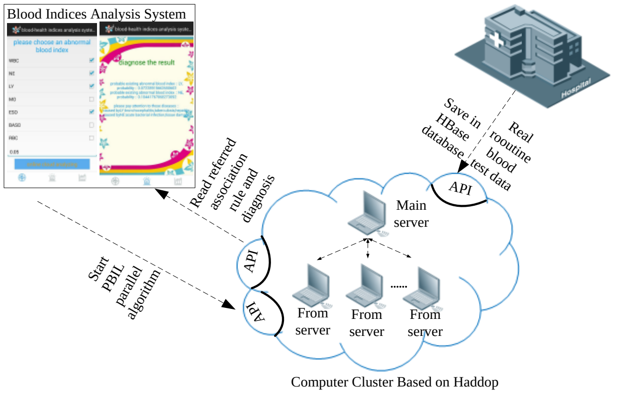

Blood Indices Analysis System
blood-health indices analysis syste...
please choose an abnormal blood index
WBC
NE
LY
MO
ESO
BASO
RBC
0.05
online cloud analyzing
diagnose the result
Hospital
Save in HBase database
Real rooutine blood test data
API
Read referred association rule and diagnosis
API
Main server
Start PBIL parallel algorithm
API
From server
From server
From server
Computer Cluster Based on Haddop

Figure 7. Architecture of blood indices analysis system based on cloud computing

VII. Conclusion

After analyzing the disadvantages of the traditional Apriori algorithm and the module of association rule mining based on GA in operation time and search quality, this paper puts forward a new module of association rule mining based on PBIL and proves that the applied algorithm performed better at searching the frequent item set comparing to Apriori and GA algorithms. Moreover, this paper designs a new parallel algorithm of association rule mining based on PBIL and architecture of the blood-health indices analysis system based on cloud computing which makes good example of the practical application of the algorithm.

References

[1] Jiawei Han. Data Mining: concept and technology [M]. Beijing: China Machine Press, 2004:137-147.J.

[2] Xiaomin Di. Research on Mining Common Risk Factors of Multi-diseases and Predicting Disease [D]. Taiyuan University of Technology, 2013.

[3] Yin Li, Changxiu Cao, Jianghong Ren, etc. Application of General Algorithm (GGA) in the Improvement of Apriori Algorithm [J]. Computer and Modernization, 2004(11):1-3.

[4] Shiwei Chen. Research on Association Rule Mining Based on Interest and Genetic Algorithm [D]. Zhejiang University, 2012.

[5] Donghao Chen, Hongwei Li, Tieying Zhang, etc. Application of Improved PSO Algorithm in Spatial Association Rule Mining [J]. Science of Surveying and Mapping, 2016, 41(2):168-172.

[6] Lämmel R. Google's MapReduce programming model — Revisited [J]. Science of Computer Programming, 2008, 70(1):1-30.

[7] Sheng Zhang. An Apriori—based Algorithm of Association Rules based on Cloud Computing [J]. Communications Technology, 2011, 44(6):141-143.

[8] Zhengchan Rao, Nianbo Fan. A review of associative rule mining Apriori algorithm[J]. Computer Era, 2012(9):11-13.

[9] Guoyan Xu, Yuqing Shi. Application of Genetic Algorithm in Association Rule Mining[j] Computer engineer, 2002, 28(7):122-124.

[10] Zhang Q. On Stability of Fixed Points of Limit Models of Univariate Marginal Distribution Algorithm and Factorized Distribution Algorithm [J]. IEEE Transactions on Evolutionary Computation,2004,8(1):80-93.

[11] Shude Zhou, Zenqi Sun. A Survey on Estimation of Distribution Algorithm [J]. Acta Automatica Sinica, 2007, 33(2):113-124.

[12] H. Muhlenbein, T. Mahnig. Convergence theory and application of the factorized distribution algorithm [J]. Comput. Inf. Technol. 1999,7(1):19–32.

[13] Qin Y J, Sun J S, Wang B Y. The differences of the blood routine indices in patients with fatty liver and non-fatty liver[J]. Journal of ClinicalHepatology,2010.

[14] Qiang Xu, Zhenjiang Wang. Practice of Cloud-computing Application Developing. Beijing: China Machine Press, 2012:64-67.

Distributed Computing System Based on Microprocessor Cluster for Wearable Devices

Xin Liu
College of Information Science and Engineering Ocean
University of China
Qingdao, China
E-mail: liuxinouc@126.com

Zhiqiang Wei*
College of Information Science and Engineering Ocean
University of China
Qingdao, China
*The corresponding author
E-mail: liuxinouc@126.com

***Abstract*—Wearable equipment in recent years has been rapid development. But the hardware manufacturing complexity and the high cost is a real problem. This paper introduces a microprocessor cluster with both hardware design principle and related distributed software design methods. This cluster has the characteristics of low cost, high reliability, flexible hardware and software system structure, low power consumption, simple equipment manufacturing process and so on, especially suitable for wearable equipment. This article discusses the hardware and software design methods in detail, as well as the complete process of the across-node communication module. In order to verify the principle of the design, we created a prototype test machine which consists of an ARM Cortex-M4 core microprocessor and 10 ARM Cortex-M0 core microprocessors through the UART serial interconnection to form a star network and carried out an experimental about the ECG feature extraction operation. Experimental results show that the performance of the cluster can be compared with a Cortex-A7 high-performance embedded processor, but the microprocessor cluster system is less expensive and has a superior cost-effective.**

Keywords-Microprocessor; Cluster; Distributed software; Wearable equipment; ECG feature extraction

I. INTRODUCTION

Wearable equipment in recent years has been rapid development. Whether in the consumer electronics market or the traditional sports apparel market, a variety of wearable devices continues to emerge. Such as an electronic wristband capable of detecting an exercise amount, a smart wristwatch capable of detecting blood oxygen and pulse rate, a sports vest with a heart rate detecting function, and a running shoe capable of detecting the sole pressure. These devices through a long time to detect human life parameters, access to traditional medical instruments can not be collected for a long time continuous data. By digging deeper into the data, the researchers found a series of hidden data features and were able to predict the health of the user for a long time.

Researchers have put forward a variety of solutions for wearable devices. On the one hand, the device is made thinner, lower power consumption, longer standby time, the sensor cable connection between the wireless connection is replaced. On the other hand, the computing power of the device is stronger, and some algorithms that need to be calculated by means of the server, after optimization, can run directly on the device.

There are many researchers working on computational methods. Through a variety of pretreatment techniques, the researchers improved the accuracy of machine learning classification to a practical level. Shereena Shaji et al. [1] increased the accuracy of certain motion recognition to 96.66%. Chin-Teng Lin et al. [2] designed an algorithm that can run on a mobile phone and detect ECG atrial fibrillation in real-time. Shigenori Shirouzu et al. [3] used a wearable device to study the relationship between electrocardiogram and sleep quality. However, the ability to enhance the software is very easy to encounter the ceiling. When some algorithms are further simplified, their accuracy will also decrease. This is an unquestionable fact.

To improve the hardware, for example, the traditional ECG acquisition requires a signal cable up to 10 cables, carrying these wires is very inconvenient for patients. Geng Yang et al. [4] proposed a sensor-based data aggregation method based on the serial bus, which solves the multi-point ECG signal acquisition problem with a single cable. In addition, how to reduce the volume, improve the integration of researchers is also concerned about the direction. G. Kavya and V. ThulasiBai [5] used an Altera FPGA chip to simulate a dual-processor dedicated system and used a parallel computation method to process the ECG signal acquisition and analysis. Jia-Hua Hong et al. [6] designed two special-purpose integrated circuit chips. One of the wireless sensor chip power consumption is very low, access to a very long battery standby time. Another receiver chip is integrated DSP processing heartbeat detection and signal classification. Shih-Lun Chen [7] designed a dedicated chip that uses less gate count and chip area to achieve higher performance.

Based on the conclusions of the researchers, we can see that highly integrated custom chips have an overwhelming advantage in terms of power consumption and performance. However, the cost of custom chips is extremely expensive. Although the traditional general-purpose microprocessors, relatively large size, and power consumption, but the technology is mature, the risk is small, and has a wealth of software and hardware resources, can reduce R & D risk. The scheme proposed in this paper is different from the traditional single-chip solution: the data processing and analysis functions are distributed to dozens of single-chip microprocessors to form a distributed microprocessor cluster. As long as the appropriate software design, the cluster can play a comparable performance of highly integrated microprocessors, and software development will not be significant difficulties.

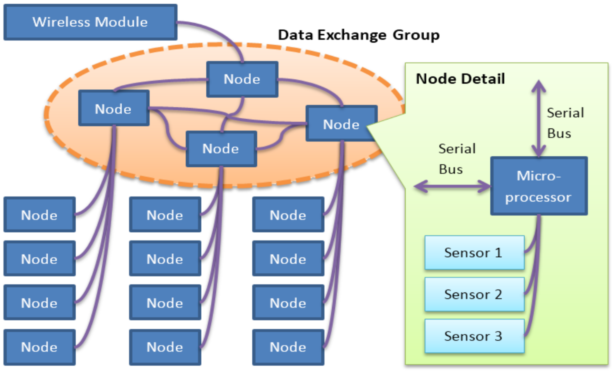


Figure 1. Block diagram of microprocessor cluster.

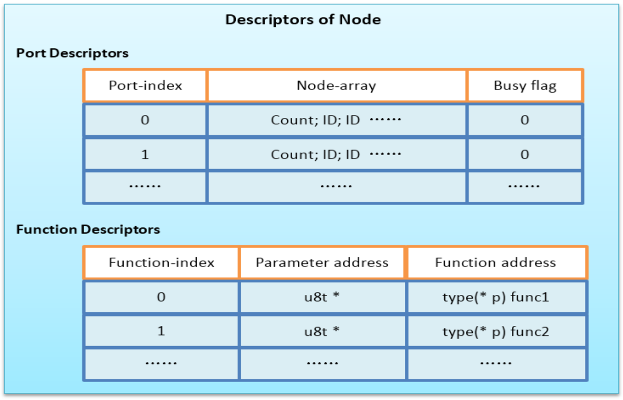
II. BLOCK DESIGN OF THE MICROPROCESSOR CLUSTER

A. *Hardware Block*

The hardware structure of the microprocessor cluster as shown in Fig .1. A cluster consists of many nodes. The core of each node device is a microprocessor. At least one UART serial port is reserved for each processor. The other ports can be used to connect a variety of sensor devices. There are also nodes located in the "Data Exchange Group", which participate in the calculation in addition to data exchange, GPIO and AD ports can be connected to a number of sensors. Data Exchange Group node compared with ordinary nodes, it needs to have a lot of UART serial port to form a Mini network. If there are enough serial ports on the chip, it is best to have a direct communication link between every two nodes. Otherwise, some inter-chip communication to be relayed through the intermediate nodes will cause performance degradation. For the microprocessor communication overhead is not negligible.

B. *Software Design*

Compared with the difficulty of hardware design, cluster software design is much more difficult. The same is true for clusters of supercomputers made up of rack-mounted servers, which are far more complex than the average desktop software.



Descriptors of Node

Port Descriptors

Port-index	Node-array	Busy flag
0	Count; ID; ID ······	0
1	Count; ID; ID ······	0
······	······	······

Function Descriptors

Function-index	Parameter address	Function address
0	u8t *	type(* p) func1
1	u8t *	type(* p) func2
······	······	······

Figure 2. Node data model.

TABLE I. PORT DESCRIPTION FORMAT SPECIFICATION.

Field name	Field format	Remarks
Port-index	Single-byte data, the range of 0 ~ 255	
Node-array	The Node-ID queues which can arrive from this port. It fixed the length of 15 bytes. Byte 1: the number of nodes n, or, queue length. Bytes 2~15: an array of node ID, only the front “queue length” IDs is available.	Because microprocessors have no more than 14 serial ports, thus reserved 14 queue length is enough.
Busy flag	Using a byte to indicates the Busy State	1 means busy and 0 means available

Simply put, the cluster nodes in the device is divided into two categories: communication nodes and computing nodes. On a server cluster, communication functions are handled by multi-layer network switches and load balancing machines, and computing functions are handled by the rack server. In this paper, the microprocessor cluster, each microprocessor has both communication nodes and computing nodes of the two functions. This paper presents a common data model to manage the functions of these two nodes, as shown in Fig .2.

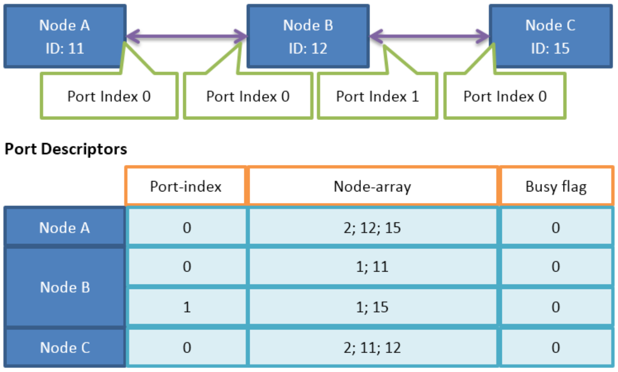
Communication is relatively independent, we first introduce the communication function. The communication function module does not generate new data and does not initiate data transmission. The new data transmission process is initiated by the calculation function module or the timer task initiative.

A Port Descriptor binds a port together. First, for each node in the cluster an ID should be assigned. Note that the ID 0 is called "the Debug Node", dedicated to using for debugging and testing the whole cluster, we will discuss the details later. Node-ID starting from 1. In general, the cluster system can't possess more than 254 microprocessors, so with 1 byte of data to store the Node-ID is enough. The value 255 is used to represent "invalid Node-ID", in some cases it may be used. Port Descriptor indicate the port linked to which nodes, and whether the port was occupied by a Communications Task, in other words, “a Busy State”. The Descriptor is shown in Table I.

So, how to deal with data relay? For example, in the simple example shown in Fig .3, the Node A send a datagram to the Node C through the Node B.

The effect of the Port Descriptors is the same as "phone book". It is stored in RAM and the microprocessor may access it at any time. When the microprocessor receives a datagram, it must first check the head of the datagram to see if the destination address is the current node. If the destination is this node, then deal with it. If not, then check the "phone book" again and select a port to send the datagram out. In the head of the datagram, therefore, requires a specific format to indicate the destination Node-ID and other information. The format of the datagram is shown in Table II.

Because the hardware connection between microprocessors has a high reliability, we need no redundancy check mechanism. We should save the communication flow. Datagram header information specified that on which function process the data, and transmit the results to which node which function. This mechanism separates the function parameters and the function returns, which can realize good flexibility, such as data collection, data preprocessing, data compression, and so on. After processing the input data, a function does not need to return the results to the source Node but comply with the Node-ID-return sent to the destination node directly. Thus avoid the unnecessary reciprocating data transmission.



	Port-index	Node-array	Busy flag
Node A	0	2; 12; 15	0
Node B	0	1; 11	0
	1	1; 15	0
Node C	0	2; 11; 12	0

Figure 3. Examples of Port Descriptors.

TABLE II. THE FORMAT OF THE DATAGRAM.

Field name	Field format	Remarks
Node-ID	Single-byte data, the range of 0~254	Destination Node-ID
Function-index	Single-byte data, the range of 0~255	Indicate which function is responsible for this data processing
Node-ID-return	Single-byte data, the range of 0~255	The Node-ID of the result receiving node. 255 means "No need to deal with the return value"
Function-index -return	Single-byte data, the range of 0~255	The function index in the receiving node which is used to handle the returned data.
Data-length	Double-byte, the range of 0~4096	The value of Data-length can be 0. As the microprocessor's RAM is very small, we limit the value up to 4096, to avoid memory overflow.
Data-data	Binary queue	Length of Data-length byte streams

As mentioned before, we called the ID 0 Node the Debug Node. When we debugging on a Node in the cluster, we can create a datagram and set its Node-ID-return value to 0 and send the datagram to the Test Node. Then, when the Test Node completes the calculation, the results will be sent to the Debug node automatically. We can install a special port in the Debug Node, such as a USB-to-Serial Bridge, to forward the datagram to PC, so that we can monitor the results easily.

C. *Distributed Computation*

When the above-mentioned communication system is constructed, the design of the calculation method becomes very simple. In practical engineering, we must first consider two issues:

1, the parameters of which nodes?

2, in which node function?

In general, the parameters are sent from a sensor node, and the function is distributed in many processors. We need a series of strategies to ensure that the parameters can be passed to the appropriate function node.

To illustrate this problem, we designed an example of the need for high-intensity computing - ECG Feature Extraction cluster. It consists of a Nuvoton M472 microprocessor and 10 Nuvoton NANO120LD3NA microprocessors (nano120) form. M472 does not participate in Feature Extraction calculation, it is only responsible for calculating the results of a summary upload. It has six UARTs and one USB port. USB interface for communication with the host computer. One of the six serial ports is used for debugging and the remaining five serial ports are connected to the serial ports of five nano120 processors (Node 1 to Node 5). Each nano120 is also connected to another nano120 (Node 6 to Node 10). ECG sampling chip Neurosky BMD100 connected to one of the nano120 (Node 10) serial port. The ECG signal comes from an ECG signal generator. The hardware structure of the whole verification system is shown in Fig .4.

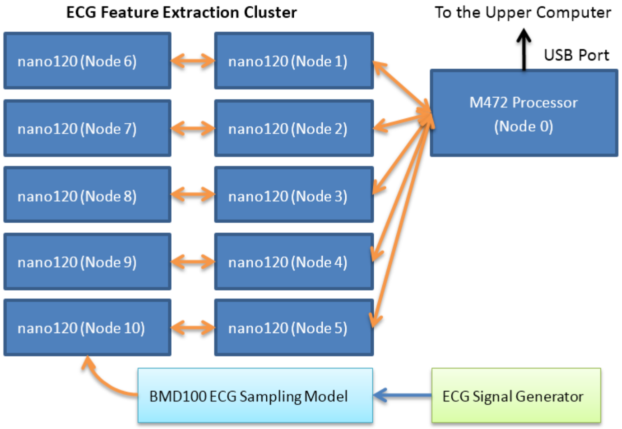


Figure 4. Block diagram of the ECG Feature Extraction Cluster.

Node 10 is a parameter node. Node 1 to Node 9 is function nodes. M472 is the data summary node. The serial baud rate between Node 10 and BMD100 is 57600. While the microprocessor is used at 115200 baud rate.

Each function node at the beginning of the calculation, the first statement to the M472 himself in a busy state, after the completion of the function calculation, and then notify the M472 to cancel their busy state. The Node 10 node first asks the M472, "Who will process the data", and sends the parameter data to the designated node after obtaining the reply from M472.

The interaction timing of each node is shown in Fig .5. Node 10 prepares a data window queue, which has three windows, in order to fill in the received ECG data. The data length of each window is two heartbeat cycles plus 120 sampling points (about 0.5 seconds of data), and the window length will vary depending on the heart rate. This length is to ensure that there are two complete ECG waveforms in each window. These three data windows are followed by a delay of one heartbeat cycle distance. When a window fills up the data, Node 10 initiates a communication process to send the window data to an idle nano120 processor. As long as the remaining 9 nano120 are not in a busy state, the whole system will not miss the ECG data. When the window data transmission is complete, the window is cleared and moved to the end of the windowing queue. Since Node 10 detects the QRS Complex and the task of allocating data to the window is heavy, it no longer assumes the Feature Extraction calculation task.

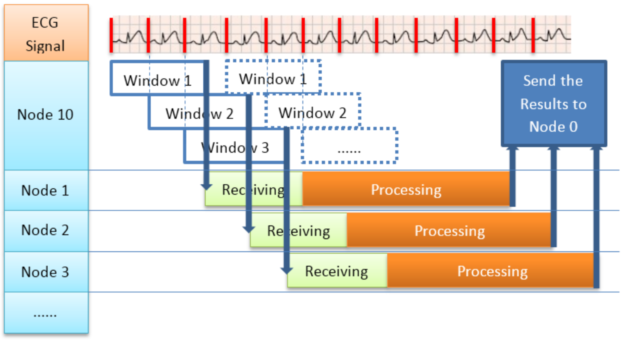


Figure 5. System Collaboration Diagram.

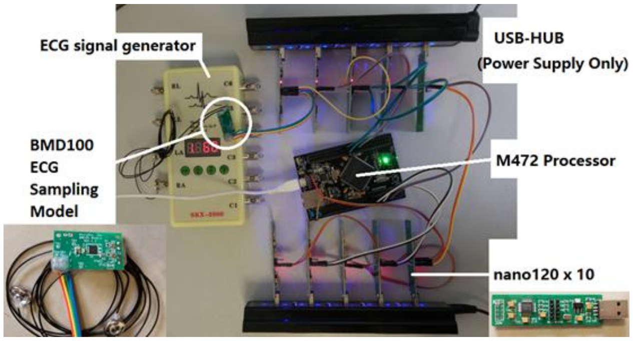


Figure 6. Hardware of the ECG Feature Extraction Cluster.

III. EVALUATION OF DISTRIBUTED COMPUTING PERFORMANCE

The completed ECG Feature Extraction cluster hardware is shown in Fig .6. We have statistics on these 9 nano120 processors. The Receiving process is relatively fixed, in the 68 ~ 75ms range, which is ECG signal generator output signal of the heart rate is fixed. When the communication baud rate is 115200, the time required to receive 620 bytes, in theory, is about 65ms. Most of the processing time-consuming and separate tests of the situation is similar to individual cases, the processing time-consuming process up to 6530ms, this is because Node 0 is sometimes communicating with the host computer, affecting the data transfer. The task assignment of the microprocessor cluster is shown in Fig .7. The numbers on the Timeline bar in the figure indicate the order of the tasks. If more than one nano120 is idle at the same time, the M472 preferentially assigns the task to a smaller number of processors.

In the experiments, we found an interesting phenomenon, the Node 8 and Node 9 has never been assigned for tasks, and Node 7 had less opportunity to be assigned for. This shown that a Cluster with 8 nano120 microprocessors was powerful enough to cope with the current computational tasks. We obtained the calculation result after the heart beat for 3 to 5 seconds, for ECG automatic diagnosis application such a delay is acceptable.

Feature Extraction is a computationally intensive process that involves filtering the signal several times and repeatedly scanning to determine the location of each wave group boundary and extremum. This process takes hundreds of milliseconds on a PC or smartphone. We tested a Spreadtrum SC9830A processor (based on Java) on a cell phone with an ARM Cortex-A7 core at 1.5 GHz and an average time of 112 ms for Feature Extraction for single-channel ECG lead data. We then ran the same test on a nano120 microprocessor (based on a C program), which took about 20 times more time. The test results are shown in Table III and Fig .8.

Compared with the SC9830A processor, nano120 microprocessor computing time is almost 20 times the former. But it's the bulk price of less than 1 US dollars, known as one dollar computer. Nano120 integrated RAM, Flash, does not require external expansion memory. ECG Feature Extraction cluster system hardware costs only 12 dollars, much cheaper than the SC9830A processor, but also to complete the same computing tasks.

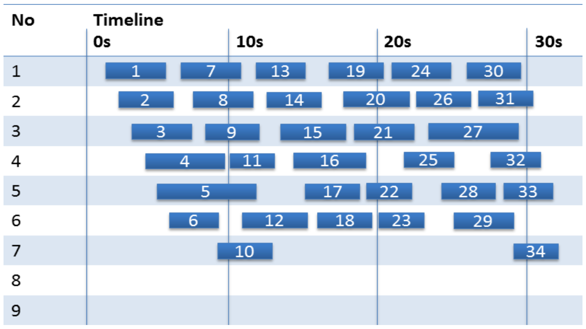


Figure 7. Task Allocation on Microprocessor Cluster.

TABLE III. SPEED COMPARISON OF THE FEATURE EXTRACTION PROGRAM ON THE SC9830A AND NANO120

No.	Consumed time (MS)	
	SC9830A	***nano120***
1	109	2398
2	110	2087
3	122	2256
4	112	2443
5	111	2299
6	108	2573
7	112	2019
8	113	2221
9	109	2320
10	109	2702
11	120	2249
12	113	2501
Mean:	112.33	2339

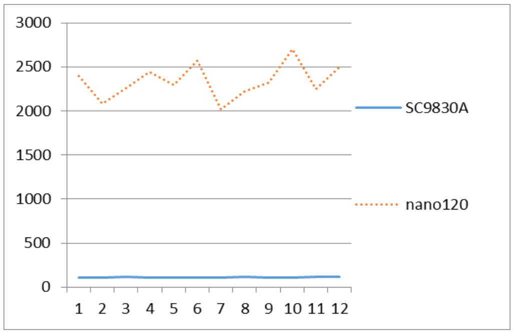


Figure 8. Speed Comparison of the Feature Extraction Program on the SC9830A and nano120.

IV. CONCLUSION AND FUTURE CONSIDERATIONS

As a popular saying goes, two heads are better than one. The same is the case with microprocessors. Described in this article the microprocessor cluster has good price-performance ratio. Its price is low and only need lower production conditions. In the aspect of software development, in order to achieve higher performance, you need to follow certain design rules. A well-designed distributed computing system can solve the problems with quite complex computation.

Outlook to the future, when wearable devices entered into the people's daily life, you can't find a part with the name "mainboard" on these devices. Every sensor node is a microprocessor, they communicate with each other via the textile fiber cable or wireless networks. With low manufacturing cost, small enough volume and long service life, the wearable devices can completely combine with clothing. Just like a mobile phone has cameras today, clothing in the future will own intelligence.

ACKNOWLEDGMENT

This work is supported by Major Projects of Shandong (No.2015ZDZX05002) and Key Research and Development Project of Shandong (No.2016ZDJS09A01)

REFERENCES

[1] Shereena Shaji, Maneesha Vinodini Ramesh and Vrindha N. Menon, “Real-Time Processing and Analysis for Activity Classification to Enhance Wearable Wireless ECG”, Proceedings of the Second International Conference on Computer and Communication Technologies. Springer India, 2016, pp.21-35

[2] Chin-Teng Lin, Kuan-Cheng Chang, Chun-Ling Lin, Chia-Cheng Chiang, Shao-Wei Lu, Shih-Sheng Chang, Bor-Shyh Lin, Hsin-Yueh Liang, Ray-Jade Chen, Yuan-Teh Lee and Li-Wei Ko, “An Intelligent Telecardiology System Using a Wearable and Wireless ECG to Detect Atrial Fibrillation”, IEEE Transactions on Information Technology in Biomedicine, Vol.14, No.3, 2010, pp. 726-733

[3] Shigenori Shirouzu, Yumeka Seno, Ken Tobioka, Tomoko Yagi, Toshiharu Takahashi, Mitsuo Sasaki and Hisanobu Sugano, “For Children’s Sleep Assessment: Can we trace the change of sleep depth based on ECG data measured at their respective home with a wearable device?”, Ieee-Embs International Conference on Biomedical and Health Informatics IEEE, 2016, pp.208-211

[4] Geng Yang, Jian Chen, Ying Cao, Hannu Tenhunen, and Li-Rong Zheng, “A Novel Wearable ECG Monitoring System Based on Active-Cable and Intelligent Electrodes”, International Conference on E-Health Networking, Applications and Services, 2008. Healthcom IEEE, 2008, pp.156-159

[5] G.Kavya and V.ThulasiBai, “Wearable advanced single chip ECG telemonitoring system using SoPC”, Ieice Electronics Express, Vol.11, No.6, 2014, pp.1-10

[6] Jia-Hua Hong, Shuenn-Yuh Lee, Member, IEEE, Ming-Chun Liang, Cheng-Han Hsieh, and Shih-Yu Chang Chien, “A Wireless ECG Acquisition and Classification System for Body Sensor Networks”, Engineering in Medicine and Biology Society IEEE, 2013, pp.5183-5186

[7] Shih-Lun Chen, Min-Chun Tuan, Tsun-Kuang Chi, and Tin-Lan Lin , “VLSI architecture of lossless ECG compression design based on fuzzy decision and optimisation method for wearable devices”, Electronics Letters, Vol.51, No.18, 2015, pp.1409-1411

[8] Joseph J. Oresko, Zhanpeng Jin, Jun Cheng, Shimeng Huang, Yuwen Sun, Heather Duschl, and Allen C. Cheng, “A Wearable Smartphone-Based Platform for Real-Time Cardiovascular Disease Detection Via Electrocardiogram Processing”, IEEE Transactions on Information Technology in Biomedicine A Publication of the IEEE Engineering in Medicine & Biology Society, Vol.14, No.3, 2010, pp.734-740

[9] Jian Kang Wu, L. Dong, X. Chen, Wee Soon Yeoh, and Isaac Pek, “Ambulatory Examination and Management of CVD Patients”, IEEE Biomedical Circuits and Systems Conference IEEE, 2007, pp. 199-202

[10] Kai Yang, Zhuan He, Wendi Yang, Qi Tang, Dongmei Li, and Zhihua Wang, “Heart Sound Denoising Using Computational Auditory Scene Analysis for a Wearable Stethoscope”, IEEE, International Midwest Symposium on Circuits and Systems IEEE, 2013, pp.1220-1223

[11] Taegyun Jeon, Byoungho Kim, Moongu Jeon and Byung-Geun Lee, “Implementation of aportable device for real-time ECG signal analysis”, BioMedical Engineering OnLine, Vol.13, No.1, 2014, pp. 160

[12] Shuang Zhu, Jingyi Song, Balaji Chellappa, Ali Enteshari, Tuo Shan, Mengxun He, and Yun Chiu, “A Smart ECG Sensor with In-Situ Adaptive Motion-Artifact Compensation for Dry-Contact Wearable Healthcare Devices”, International Symposium on Quality Electronic Design, 2016, pp.450-455

[13] G. Tortora, R. Fontana, S. Argiolas, M. Vatteroni, P. Dario, M.G. Trivella, “A dynamic control algorithm based on physiological parameters and wearable interfaces for adaptive ventricular assist devices”, International Conference of the IEEE Engineering in Medicine & Biology Society, 2015, 4954

[14] Chao-Ting Chu, Huann-Keng Chiang, and Jian-Jie Hung, “Dynamic Heart Rate Monitors Algorithm for Reflection Green Light Wearable Device”, International Conference on Intelligent Informatics and Biomedical Sciences IEEE, 2015, pp.438-445

[15] Sungmook Jung, Jongsu Lee, Taeghwan Hyeon, Minbaek Lee, and Dae-Hyeong Kim, "Fabric-Based Integrated Energy Devices for Wearable Activity Monitors", Advanced Materials, Vol.26, No.36, 2014, pp.6329-6334

[16] Ya-Li Zheng, Bryan P. Yan, Yuan-Ting Zhang, and Carmen C. Y. Poon, "An Armband Wearable Device for Overnight and Cuff-Less Blood Pressure Measurement", IEEE transactions on bio-medical engineering, Vol.61, No.7, 2014, pp.2179-2186

[17] Min Chen, Yin Zhang, Yong Li, Mohammad Mehedi Hassan, and Atif Alamri, "AIWAC- Affective Interaction Through Wearable Computing and Cloud Technology", IEEE Wireless Communications, Vol.22, No.1, 2015, pp.20-27

[18] Gianmarco Angius, Luigi Raffo, "Cardiovascular Disease and Sleep Apnoea: a Wearable Device for PPG Acquisition and Research Aims", Computing in Cardiology, Vol.39, 2012; pp.513-516

[19] Biyi Fang, Nicholas D. Lane, Mi Zhang, Aidan Boran, Fahim Kawsar, "BodyScan : Enabling Radio-based Sensing on Wearable Devices for Contactless Activity and Vital Sign Monitoring", The International Conference, 2016, pp.97-110

[20] A.J. Cook, G.D. Gargiulo, T. Lehmann and T.J. Hamilton, "Open platform, eight-channel, portable bio-potential and activity data logger for wearable medical device development", Electronics Letters, Vol.51, No.21, 2015, pp. 1641-1643

Implementing Business-to-Customer, IT Outsourcing and Workflow Management System to Exploit China Education Market for Durham University

Jiacong Zhao
Department of Information Management
Dalian Neusoft University of Information
Dalian, China
E-mail: zhaojiacong@neusoft.edu.cn

Jingshu Wang *
Department of Information Management
Dalian Neusoft University of Information
Dalian, China
E-mail: wangjingshu@neusoft.edu.cn
*The corresponding author

Chuanlin Huang
Department of Information Management
Dalian Neusoft University of Information
Dalian, China
E-mail: huangchuanlin@neusoft.edu.cn

Chen Qian
Department of Information Management
Dalian Neusoft University of Information
Dalian, China
E-mail: Qianchen14@nou.com.cn

***Abstract*—Currently, education has become a business sector that provides service to students. This service is developing standards more similar to consumer goods marketing. As a higher education organization, Durham University aims to top education and research across all disciplines in 2020 in UK, which needs the support of the most able and motivated students, academic staff and sufficient money. During the past two years China was the top international students sending country to UK, who pay highly tuition fees. Chinese shows high academic capabilities in variety academic field as well. In this situation, DU should enhance its current strategies to exploit this most talent market with implementing Business-to-Cusiness (B2C), IT outsourcing (ITO) and Workflow Management System (WfMS). These initiatives, significantly achieving customer satisfaction (CS), long-term development, flexible risk management, cost saving, short develop timescales and effective organizational management. Importance-Performance Analysis tool and decision value theory will be introduced to evaluate these initiatives.**

Keywords-B2C; ITO; WfMS; Customer satisfaction; Cost saving

I. INTRODUCTION

Currently, education has become a business sector that provides service to students [1]. This service is developing standards more similar to consumer goods marketing [2]. Durham University (DU) owns high academic reputation in human sciences and aims to top education and research across all disciplines in 2020 in UK [3,4], which needs the support of talented students, staff and sufficient money. China was the top overseas students sending country to UK Higher Education (HE) in 2013/14 and 2012/13, which separately shares 20.28% and 19.84% of UK's international students [5]. Additionally, Chinese currently show highly academic abilities. Hence, DU typically places its market in China and promote e-business on attracting the most able and motivated students and academic staff. Companies likely to succeed in e-business concentrate on linking e-business knowledge to their core business, enabling technology, gaining customer satisfaction (CS) and maintaining online operations [6]. Hence, DU will focus on achieving CS, overcoming international business barriers, saving cost, enhancing management capabilities. Business-to-customer (B2C) [7], IT-outsourcing (ITO) [8] and Workflow Management Systems (WfMS) [9] are introduced to deal with these issues. This report prioritizes these initiatives as B2C, ITO and WfMS.

II. DU

A. Student Market Organization

DU operates recruitment by cooperating academic staff of each discipline with marketing office [10]. Fig .1 shows the scope of the organization. DU finance experienced gradually increases in terms of net income for the past five years (Fig.2). However, its finance takes no advantages compared with peer universities in terms of rank and reputation (Tab. I). Fig .3 [11] shows DU's 2014 income components. Tuition fees is clearly one of the main kinds of income to DU. Compared with Home students, Chinese students pays about four times of fees, such as Table II .

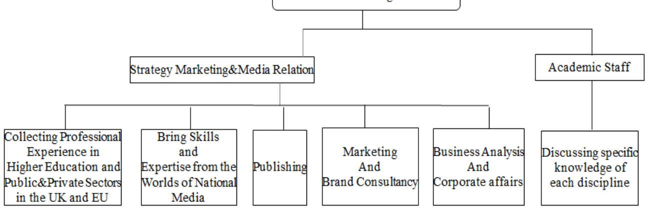

Student Market Organization
Strategy Marketing&Media Relation
Academic Staff
Collecting Professional Experience in Higher Education and Public&Private Sectors in the UK and EU
Bring Skills and Expertise from the Worlds of National Media
Publishing
Marketing And Brand Consultancy
Business Analysis And Corporate affairs
Discussing specific knowledge of each discipline

Figure 1. DU Student Market Organization

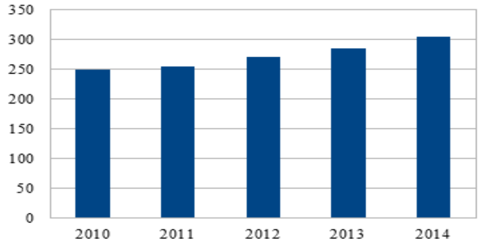


Figure 2. The Net Income of Durham University (million) for Five Years

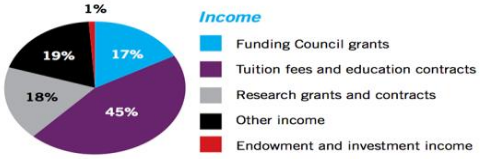


Figure 3. Income Component of Durham University

TABLE I. NET INCOME AND ASSET OF PICKED UNIVERSITIES AT 2016

2016	Net Income(m)	Net Assets(m)
University of Durham	283.34	287.89
University of Manchester	826.97	826.54
University College London	937.24	811.7
Imperil College London	822	1002

TABLE II. TUITION FEES FOR ACADEMIC YEARS 13/14 AND 14/15

	13/14	13/14	14/15	14/15
Nationality	Home	Chinese	Home	Chinese
Classroom(PGR)	£3,900	£13,300	£3,996	£14,000
Laboratory Based(PGR)	£3,900	£17,000	£3,996	£17,900
Premium Classroom(PGR)	£3,900	£13,300	£3,996	£14,000

B. IT System (ITS) and Website Quality

DU's ITS first introduced in 2000 [10] and ran about 15 years, which provides tools and applications that support campus-wide business and academic applications. The ITS's last year's cost is around £3,580,000 which mainly includes IT infrastructure, maintenance, depreciation and staff cost. This ITS is based on the theory of Blackboard Systems (BS) [10] (Fig .4). This enables DU to figures its basic components as Fig .5. Knowledge Sources (KSs) contain the problem-solving knowledge and each KS works independently. Blackboard is a global structure for all KSs. Control component directs the problem solving process by allowing KSs to respond to changes on the blackboard database [10]. Robust database system and enjoys edit flexibility.

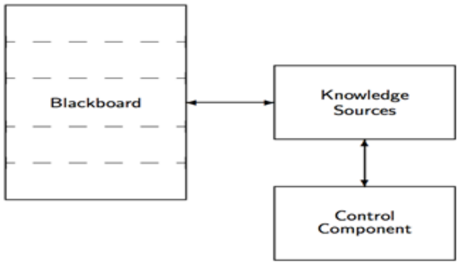


Figure 4. Basic Components of the Blackboard Model

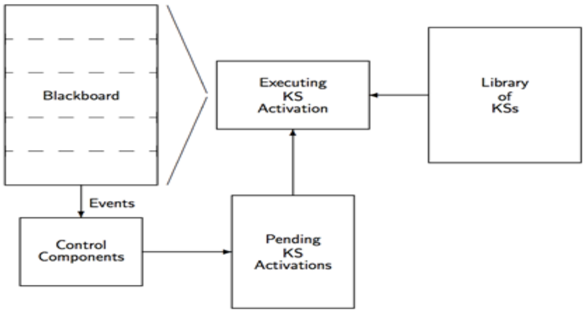


Figure 5. Basic Components for Durham University's IT System

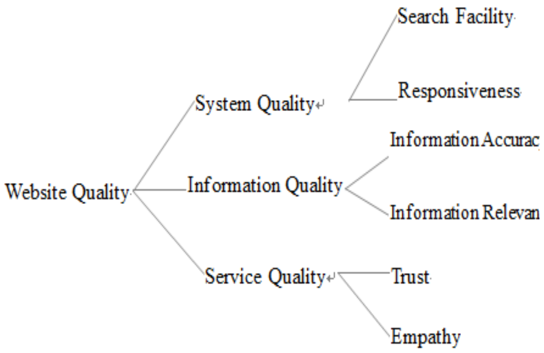


Figure 6. Website Quality Components

Website quality (Fig .6) should consider both system-oriented and service-oriented quality [12]. System-oriented quality mainly refers to search facility and responsiveness. Search facility reflects whether a tool or structure actually helps a website user to find perceived information [13]. DU includes a search engine in its ITS to mitigates the difficulty to find all the information in a specific subject. Additionally, tools like menus, frames and image maps are processed to avoid navigation problem [14]. DU accesses responsiveness by shorting search and load time [15]. Hence, DU runs well system-oriented quality. However, the lack of customer knowledge impedes service-oriented quality.

Therefore, DU's ITS can well support its e-business. The weaknesses are customer-related issues, limited financial budget and highly human-related recruiting organization. With adopting suitable strategies, current capabilities and resources can support DU exploits China market.

III. StylingStrategic E-business Initiatives

DU recruits Chinese through two traditional ways: exchanging students and staff with local universities and dealing with personal online application. Although it admitted enough people, it not means that these people are qualified as the university expectation. Poor business performance and repeat patterns of existing behavior are symptoms of failure [15]. Thus, DU updates its current e-business behavior by conducting B2C, ITO and WfMS.

A. B2C

B2C refers to companies and their customers perform online commerce via Internet-based technologies [16]. B2C concentrates on developing knowledge relevant to their core business and considers changes that may occur in customer relationships and create response solutions [6]. Additionally, a firm's effectiveness on fulfilling orders in B2C transactions is a significant determinant of customer satisfaction (CS) [17]. It can be seen that B2C combines of strong customer focus. Increasing CS increases customer loyalty [18]. However, there are also impediments for B2C adoption. For example, compared with e-business, CS is less challenging in conventional business [19] where customers are immediately empowered with the required information for decision making. It indicates that delivering trustworthy information to customers are crucial in e-business. Thus, B2C should manage both system-oriented and serviced-oriented CS. System-oriented CS mainly refers to provide customer with a usably, availably and effectively website. Service-based CS especially satisfies customers through providing highly accuracy and relevance information and trustworthy service.

One of DU customer-related issues is trust which is well recognized as one of the strongest effects on e-business [20]. Based on B2C, trust can be achieved through both system-oriented and service-oriented level. DU has successfully attained highly system-oriented quality as mentioned. In service-based level, DU should offer a guest role to potential members. This service enable them share ideas with professors and students freely and access trustworthy information. This real life experience of being a member of the university improves their confidence to the university. Therefore, B2C adds significantly CS advantages to DU, which provides the wanted members more opportunities to know the university and then trust the university.

B. ITO

ITO means handing over the management of part or all of an organization’s information technology, systems and related services to a third party [11]. It is well recognized that global ITO is a lucrative alternative in capital market gains cost savings, skilled labor and short marketing development time [21]. Furthermore, it collects local data for marketing and customer analysis effectively. It proves that cost reduction, business performance improvement are the mainly motivations for outsourcing [22]. However, varieties of barriers prevent companies integrating business into global market, like time zone difference, channel conflicts, legacy system and cultures. Current study shows ITO performance is significantly related to the following three factors of organization:

1) Finance: Compared with peers, poor finance companies are more likely to conduct ITO and organization size and industry have no significant effect [21, 22].

2) Percentage of IT budget and/or types of outsourced functions: Companies outsourced less than 80% had success rate of 85%, oppositely success rate was 29% [23].

3) Types of outsource functions: Outsourcing system operations and business process gain higher levels of satisfaction than outsourcing systems management and applications development [24].

The characters of ITO is suitable to DU. Firstly, DU's finance situation takes no advantages compared with peers (Tab .II). Secondly, DU only outsources less than 40% of its marketing functions (Fig .1): marketing and brand consultancy, business analysis and cooperation affairs. Finally, in China there are numbers of high quality intermediary agents that provide DU a large pool for cooperation. Therefore, ITO contributes cost mitigation, short timescale and effectively marketing specification to DU.

C. WfMS

WfMS is a system that completely defines, manages and executes workflows through the execution of software whose order of execution is driven by a computer representation of the workflow logic [9]. Specifically, WfMS allows the user to define and design different workflows, like ad hoc, administration and production [25], for different types of business processes [27]. Furthermore, WfMS keeps tracking all processes simultaneously to control and coordinate workflow and information between participants to automate processes [9]. However, WfMS's limitation in functionality usually leads to actual features provided by the systems are not well correlated to the expectations from the users. For example, WfMSs support ad hoc workflows must provide functionality for facilitating human coordination, collaboration and co-decision, but such WfMSs cannot be used for workflows for controlling task ordering. Therefore, WfMS's contribution to the business performance based on the serious analysis of business processes of an organization and the corresponding workflows.

DU's current student recruiting system is partly automated. Specifically, only the online application process is automated, but not a cohesive and automatic

TABLE III. Importance mark for the valuable attributes and Performance mark for each initiative

Categories	Success Attributes	Mark for Importance	Mark for B2C Performance	Mark for ITO Performance	Mark for WfMS Performance
Strategic Aspect	Interrupted core business activities with website quality	5	5	3	2
	Kept traditional competitive advantages	5	4.9	2.3	4
	Improved market share track	1	3.4	4	2
	Relations enhanced competitive advantage	3	4.3	4	3
	Improved new competitors track	1	1	4.1	2.1
	Improved buyer behavior track	4	4.1	5	2
	Offered customer personalization	4	5	3	4
	Quicker timescale to market	4	3	4.9	2.2
	Good services offered by e-business	4	3.9	3.1	3
	Innovation allowed when risks are low	5	4.8	4.9	2.1
	Improved customer satisfaction	5	4.7	5	4.1
	Trained employee	2	2	5	3
	Good cost control	5	3.3	4.8	3
	Improved response to change	3	3	3	4
	Improved service quality	4	4.6	4	4.2
	Improved team work	2	3.3	2	4
	Overcome culture barriers	3	2.2	5	2
	Promote proactive culture	3	2	4.1	2.1
	Improved planning	3	3.2	3.1	3.1
	Improved administrative procedures	3	4.2	3.2	3.2
	Reduced market cost	5	3.8	4	2.2
	Leader in new technology	2	3	4.8	3.1
	Improved organizational process flexibility	5	3	2	4.2
Management-oriented Aspect	Improved capabilities understanding of technology by executives	4	4.3	2	4.4
	Support for e-business from top management	4	4.2	2.1	4.5
	Improve communication throughout the organization	4	3.1	2.2	5
	Improve communication with customers	5	4.6	4.7	5
	Improved data management	3	4.5	3.3	3.3
	Reduce paper work	1	3	3	4
	Reduce labor cost	4	3.2	4.9	3.6
	Reduce rework	2	3.2	2.2	3.2
	Improve quality of out put	1	5	3.5	4.2
	Improve ability to exchange data	3	3.1	4.8	4.3
	Improve response time to queries	4	3	4.9	4.6
	Improved forecasting and control	3	4.7	4.2	3.5

TABLE IV. ADVANTAGES AND DISADVANTAGES OF INITIATIVES

B2C		**ITO**		**WfMS**	
Advantages	**Disadvantages**	**Advantages**	**Disadvantages**	**Advantages**	**Disadvantages**
Integrating Online and Off-line business	Difficulty in measuring business performance	Gaining predictable technology budget	Critical information can be scary	Guaranteeing a concurrency system	Increasing expenditure for the increased scope
Offering personalization to customer	Highly accuracy requirement	Cost savings	Loss of a critical capability	Integrating cross-enterprise workflows	Inflexible in handing human intervention
Collecting customer knowledge	Long time development for the process	Greater access to skilled staff	Possible threat of opportunism from suppliers	Business process verification and simulation	Unpredictable scalability for the system
Managing relationship with customer	Attractive content is difficult to define	Focuses on the core products and services	Loss of flexibility	Matching workflow to organization strategy, structure	Poor communication support
Winning new customers	Relative large investment	Short timescale for development	Loss of advantages in information management	Automating resource and information management	Poor fault tolerance
Monitoring internal and competitors activities	Higher employee ability demand	Risk sharing with the technology partner	Decline performance of current employee	Coordinating and streamlining business processes	Failed components cannot be replaced easily

Recruiting system. For example, the communication Among applicants, marketing departments and academic is suffering great latency. DU's business process is highly information related and can be defined as ad hoc workflows. Therefore, WfMS guarantees a dynamic and automatic recruiting system to DU.

TABLE V. CONTRIBUTION VALUE DECISION MATRIX

Marking Rules		**Number of Attributes**			**Value for each Attributes**
		ITO	**B2C**	**WfMS**	
I=3	P=3	1	1	1	0
	P> 3	8	6	6	1
	P<3	0	2	2	-1
I>3	P=3	2	2	2	0.5
	P>3	12	15	11	1
	P<3	5	1	5	-1
Total Value	B2C	1*0+6*1+2*(-1)+2*0.5+15*1+1*(-1)=20			
	ITO	1*0+8*1+0*(-1)+2*0.5+12*1+5*(-1)=16			
	WfMS	1*0+6*1+2*(-1)+2*0.5+11*1+5*(-1)=11			

IV. EVALUATION AND PRIORITIZATION

After In e-business, a well performance strategy may show little importance to the success of e-business. This indicates that the adoption of a strategy should consider both importance and performance values. Hence, this report adopts the Importance-Performance Analysis (IPA) [28] which is a tool for evaluating marketing strategies, to analyze the performance of initiatives for the important attributes of DU. Then prioritize initiatives with decision value theory [30]. The process follows these steps:

(1).Mark picked success attributes for DU in terms of importance and performance of each initiative to attributes.

(2).Draw IPA maps

(3).Calculate contribution value of initiatives

A. Marking for the Performance and Importance

Cooperating analysis of DU and success factors summary in e-business identified from the literature [6, 7, 8, 18, 30,31, 32, 33] presents the success attributes to evaluate these initiatives. The case study of [3, 7, 8, 15, 24] marks the importance of these attributes (Table. III). The review of [2, 6, 16, 17, 18, 20, 26, 27], [11, 21, 22, 23, 24] and [25, 26, 27] separately shows the advantages and disadvantages derived from B2C, ITO and WfMS to DU (Table IV), and marks the performance of them in Table. III. Based on the importance and performance value drawing the IP maps (Fig .7, Fig .8, Fig .9).

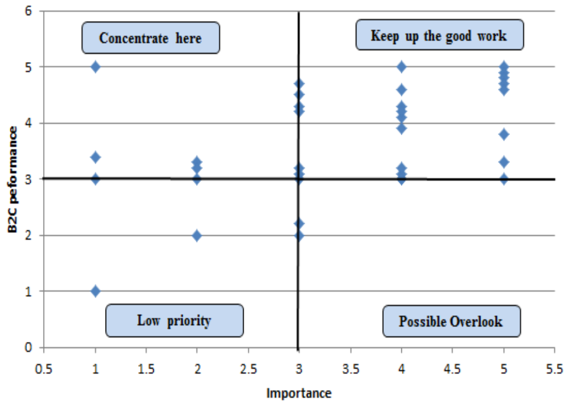


Figure 7. IP maps for B2C

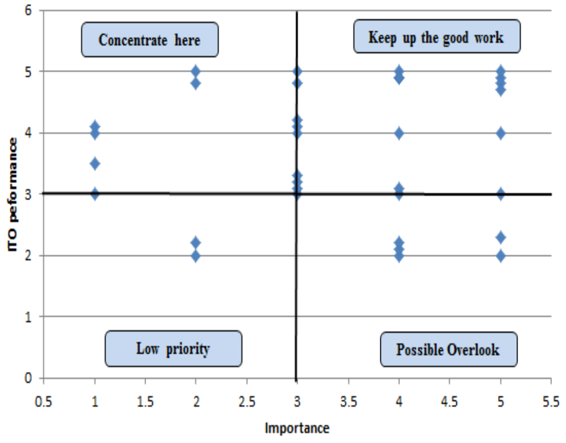


Figure 8. IP maps for ITO

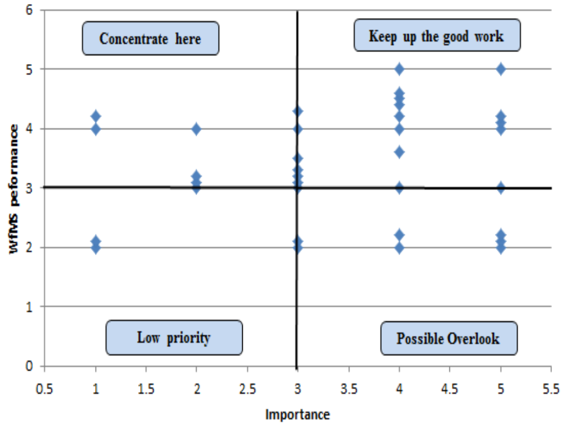


Figure 9. IP maps for WfMS

B. *IPA and Comparison of Initiatives*

IPA demonstrates each initiative's performance to the importance of DU. Interpretation of the four quadrants of IPA is illustrated:

(1).Concentrate here: bad performance for the importance.

(2).Keep up the good work: well performance for the importance.

(3).Low priority: badly performance for the unimportance.

(4).Possible overkill: well performance for the unimportance.

It shows that (1) and (2) seriously reflect the contribution of each initiative for DU's importance. However, (3) shows low salience in effecting the business and (4) even states little contribution. Hence, the comparison factors especially focus on the attributes (1) and (2). Although the IPA clearly locates these attributes, the evaluation of importance-performance of these initiatives should base on decision value theory [30]. Finishing final comparison with following steps:

(1). Calculating and marking the contribution value of these attributes with following rules (I donates value of Importance, P donates value of Performance):

If	And	Mark
I=3	P=3	0
	P>3	1
	P<3	-1
I>3	P=3	0.5
	P>3	1
	P<3	-1

(2).Calculate the contribution value for each initiative, and rank them (Table V)

The calculated total value (Table V) prioritizes these initiatives as B2C, ITO and WfMS. The attributes locations of IPAs show that B2C takes the advantages of customer relationship management, risk reduction and strategy flexibility. ITO shorts marketing development lifecycle and saves cost, WfMS extremely enhances the execution of organization. This demonstrates that DU should firstly focus on CS, then boost the productivity to China market and finally improve organizational management to support the e-business.

C. *Limitation*

The importance and performance are marked by predication that based on literatures. These mark maybe not accurate enough. Then, the contribution values are marked by widely divided range of I and P, should value them more seriously. But the outcome of the evaluation is reasonable to DU.

V. Conclusion

What E-business is widely adopted in education market enables universities to recruit the most able and motivated students and staff across the world, so as to improve the education and research level and boost financial budget. While create a successful collaboration between business strategies and supported technologies is composure for e-business adoption. Hence, firstly analyzing both the internal and external environment for DU to identify its current capabilities and resources to support its core business goals. Then picking B2C, ITO and WfMS to attain its customer relationship management, global cooperation, cost saving and motivation of internal working mechanism. The IPA and decision value theory prioritizes these initiatives as B2C, ITO and WfMS. The evaluation results also indicate that each initiative takes advantages in a typical field; e-business success should base on the cooperation of them, or proposed a comprehension initiative to model different advantages simultaneously.

References

[1] Kalakota, R. and Whinston, A. , “Electronic Commerce: A Manager's Guide”, Addison-Wesley, Reading, MA,1997.

[2] Singh, M., “E-Services and Their Role in B2C E-Commerce’, Journal of Managing Service Quality, 12:2, pp.434 – 446, 2002.

[3] Durham University Annual Report 2013

[4] Durham University Annual Report 2012

[5] https://www.hesa.ac.uk/

[6] Dubelaar, C., Sohal, A., Savic, “V. Benefits, Impediments and Critical Success Factors in B2C E-business Adoption”. Technovation 25 (11), pp.1251–1262, 2005.

[7] Love, P. E. D., Irani, Z., Standing, C., Lin, C., & Burn, J. M., “The Enigma of Evaluation: Benefits, Costs and Risks of IT in Australian Small-medium-sized Enterprises”. Information & Management, 42(7), pp.947–964, 2005.

[8] D. Phan, “E-business Development for Competitive Advantages: a case study”, Information and Management 40 (6), pp.581–590, 2003.

[9] D. Hollinsworth, “The Workflow Reference Model, Workflow Management Coalition”, TC00-1003, December,1994.

[10] Bradford, Peter, et al. "The Blackboard learning system: The be all and end all in educational instruction?." Journal of Educational Technology Systems 35.3, pp. 301-314, 2007.

[11] Willcocks, L., Fitzgerald, G. and Feeny, D. , Outsourcing IT: the strategic implications", Long Range Planning, Vol. 28 No. 5, pp.59-70, 1995.

[12] Huizingh, E., “The Content and Design of Websites: An Empirical Study”, Information & Management, Vol. 37 No. 3, pp.123-34, 2000.

[13] Levene, M., “Web dynamics, Focus Review”, Vol. 2 No. 2, pp.60-71,2001.

[14] Wan, A.H., “Opportunities to Enhance a Commercial Website”, Information & Management, Vol. 38 No. 1, pp.15-21, 2000.

[15] Thorne, M.L., “Interpreting Corporate Transformation through Failure”. Management Decision 38 (5), 2000

[16] Ranganathan, C. & Ganapathy, S., “Key Dimensions of Business-to-consumer Websites, Information and Management”, 39, pp.457–465, 2002.

[17] Shenton, J., E-Business Brings E-Fulfillment, 2002. http://www.globalmillenniamarketing.com/article_efulfillment_ecommerce_ebusiness.html.

[18] Turban, E., Lee, J., King, D. and Chung, H.M., “Electronic Commerce: A Managerial Perspective”, Prentice-Hall International Inc., Englewood Cliffs, NJ, 2000.

[19] Bhattacherjee, “Understanding information systems continuance: an expectation confirmation model”, MIS Quarterly 25(3), pp.351–370, 2001.

[20] Lin, H. F., “The Impact of Website Quality Dimensions on Customer Satisfaction in the B2C E-commerce Context”, Total quality management & Business Excellence, 18(4), pp.363–378, 2007.

[21] Pastore, M., “E-commerce Faces Logistics Nightmare”,1997.

[22] Http://cyberatlas.internet.com/markets/retailing/article/06061_19045100.html

[23] Levina, Natalia, and Jeanne W. Ross. "From the vendor's perspective: exploring the value proposition in information technology outsourcing." MIS quarterly, pp. 331-364, 2003.

[24] Lacity, M., Willcocks, L., “An empirical investigation of information technology sourcing practices: lessons from experience”. MIS Quarterly 22 (3), pp. 363– 408,1998.

[25] Grover, V., Cheon, M., Teng, J., “The effect of service quality and partnership on the outsourcing of information systems functions”. Journal of Management Information Systems,vol.12 (4), pp. 89–116,1996.

[26] D. Georgakopoulos, M. Hornick, and A. Sheth, “An overview of workflow management: from process modeling to workflow automation infrastructure,” Distributed and Parallel Databases, vol. 3, pp. 119–153, 1995

[27] ALONSO, G., AGRAWAL, D., EL ABBADI, A., AND MOHAN, C., “Functionality and limitations of current workflow management systems”. Computer Science Department, University of California at Santa Barbara, Santa Barbara, CA, 1997.

[28] Fishburn, Peter C, “Decision and value theory”. New York: Wiley, No. 10., 1964.

[29] Martilla, J.A., and James, J.C, “Importance-Performance Analysis”, Journal of Marketing (41:1), pp.77–79, 1977.

[30] Damanpour, F., “E-business E-commerce Evolution: Perspective and Strategy”, Managerial Finance 27 (7), pp.16–33, 2001.

[31] Sharma, P., “E-transformation Basics: Key to The New Economy”. Strategy and Leadership 2 (4), pp.27–31, 2000.

[32] Porter, M.E., “Strategy and The Internet”. Harvard Business Review March, pp.62–78, 2001.

[33] Hackbarth, G. and Kettinger, W.J., “Building an E-business Strategy”, Information Systems Management, Summer, pp. 78-93, 2001.

Novel Model of E-commerce Marketing Based on Big Data Analysis and Processing

Hongsheng Xu*
Henan key Laboratory for Big Data Processing &
Analytics of Electronic Commerce
Luoyang Normal University
Henan LuoYang, China
E-mail: 85660190@qq.com
*The corresponding author

Ganglong Fan
Henan key Laboratory for Big Data Processing &
Analytics of Electronic Commerce
Luoyang Normal University
Henan LuoYang, China
E-mail: 85660190@qq.com

Ke Li
Henan key Laboratory for Big Data Processing &
Analytics of Electronic Commerce
Luoyang Normal University
Henan LuoYang, China
E-mail: 85660190@qq.com

***Abstract*—The amount of information on the Internet is getting larger and larger, and the energy of the consumer and the ability to deal with information is limited. Electricity supplier enterprises in the development process to do are to use big data for personalized shopping guide. This paper analyzes the situation of the development of e-commerce industry in the background of big data, and puts forward the improvement method. The paper presents novel model of e-commerce marketing based on big data analysis and processing.**

Keywords-Big data; E-commerce; Network marketing; O2O; Massive data

I. INTRODUCTION

Big data may be heterogeneous data obtained from different physical address on the data source, and or with multi modal characteristics, is the typical representative of social media data are widely distributed on the Internet. Such as micro-blog data contains micro-blog users of gender age occupation and social networking, natural text and image, such as video and audio. For such data, even if the investigation under the same attribute of different physical space objects of the same class, cannot expect its has the same distribution characteristics. The symbols, values, time series, image, text and social network structure of a single mode, equivalence relation, symbolic and numerical characteristics of structural similarity the relationship or order relation based on object classification, can form a covering or nested grain structure.

The era of big data first determined is by data richness. The rise of social networks, a large number of UGC (Internet terminology, meaning the meaning of user generated content), audio, text information, video, pictures and other unstructured data appeared. In addition, the greater the amount of information on the Internet of things, coupled with the mobile Internet can more accurately and quickly collect user information, such as location, life information and other data. From the amount of data, has entered the era of big data, but now the hardware has been unable to keep up with the pace of data development.

Big data is more in the national economy and the use of marketing. With the development of mobile Internet and intelligent cloud technology, big data has begun to affect more and more business. That is to be able to more effectively on the supply chain, product development, online drainage guide, thereby enhancing the efficiency of the operation of the platform. Big data provides the possibility of e-commerce, and has begun to more and more applications [1]. Key to the success of e-commerce upgrade that is necessary to highlight the traditional doors and windows retail service and experience, but also the ideological front of the Internet into the brand concept, in this context, e-commerce O2O thinking emerged, big data has become a key O2O loop.

Internet era, user habits are changing, and it is only a full understanding of the user to create products in line with user expectations. Compared with traditional retail, e-commerce, the biggest advantage is that everything can be monitored and improved through the data. Through the data can be seen from where the user, how to organize the product can achieve a good conversion rate, the efficiency of advertising and so on. Every bit of change based on data analysis is a little bit more profitable.

The data is a relatively abstract concept, especially when faced with massive data easily confusing, traditional data analysis is more to show, in a simple chart or PPT form is not intuitive, since 2010 the data information map rise, provides visual effects and understanding is very good for data analysis and output, he used a combination of simple graphics will be converted to a single chart connotation rich results, greatly stimulate people's senses, the boring data and more vivid information graph is a manifestation of the further development of data visualization, the era of big data will lead to a lot of a similar method.

Electronic commerce is the first in the network business, is using the Internet data to understand each customer's needs and tendency to provide personalized products and services for them, and then quickly and easily realize the transaction and delivery of products and services to the user as possible, personalized service features, the automation of the relationship for commercial organizations to increase the income and cost, to establish and strengthen relationship with customers.

In the big data era, e-commerce in the economic activities of the operating mode from traditional management into data management mode, performance management process and economic activities of electronic commerce business enterprise data activities, data operation development of the concept of penetration in the enterprise raw material procurement, product manufacturing and product the whole process of marketing activities. The use of electricity in the process by using professional data analysis technology will be able to carry out comprehensive and prediction of certain induction on consumer habits and consumer psychology, and clever use of distribution adjustment of product market supply and demand, and it is in order to satisfy consumer demand of consumers, reduce the cost of production and sales of products. The paper presents novel model of e-commerce marketing based on big data analysis and processing.

II. Analysis on the Development of E-commerce Industry under the Background of Big Data

Big data marketing refers to the behavior of a large amount of data collected through the Internet, first of all to help advertisers find the target audience, the advertising content, time and form of anticipation and allocation, and ultimately complete the process of marketing advertising.

Big data can pass the verification and evaluation of massive data, increasing the risk of controllable line and management, timely find and solve the possible risk points, have accurate grasp for the regularity of the risk, will drive demand for more in-depth and thorough analysis of the data of financial institutions. Support business refinement management [2]. Although the bank has a lot of data to pay for water, but the Department is not cross, the data can not be integrated, big data banking model prompted the bank began to effectively use the data deposited. Big data will promote the innovation of financial institutions and service brand, do fine service, are customized to the customer, the use of data analysis and prediction of the new development model, realize the analysis of customer consumption patterns to improve customer conversion rate. Big data is bound to bring more opportunities for financial companies to update the data based business and internal management optimization.

The security and integrity of data is the core elements of the enterprise must consider the elements of e-commerce sites in accordance with the three levels to do this, the first level: the basic hardware. In data storage, data backup, most companies can do. The second level is the key data management, who can see these data. This level is the largest investment in the company, as is shown by equation (1), where p is in the management of data, a lot of data leaks may be internal reasons. There is some internal staff to see the data he should not see, or get the data he should not get [3].

$$p_0(t) = \frac{u}{\lambda + u} + \frac{\lambda}{\lambda + u} e^{-(\lambda+u)t} \tag{1}$$

This is an era of information explosion, in the face of a large number of consumer choice, the same marketing approach has been difficult to attract consumer's loyal consumption. Therefore, today's retail marketing must come up with the content and form that can really touch the hearts of consumers. We believe that to do this, we must rely on the collection of electricity supplier and store sales data as one of the large database. Through the analysis of these data, we can not only accurately understand each member's age, gender, consumer consumption, consumption frequency, also can analyze their shopping habits and preferences, in shopping malls throughout the residence time and order from.

Found that grain size model (fusion) is a natural big data super large-scale, multi modality, hybrid features, internal logic requirements for solving complex problems and granular computing under the framework of big data. After granulation, the particle pattern of each homogeneous data sample from the perspective of problems can be solved by the existing methods though however, global data on pattern discovery needs to carry on the fusion strategy under the guidance of the heterogeneous data. The particle (corresponding to different feature subspaces of the abstract concept level, and according to the different modes under the framework of data granulation results) on the issue of pattern discovery, as shown in equation (2) [4].

$$u_c(t) = \begin{cases} 0, 0 \le t \le c, \\ \frac{1}{4}(t-c)^2, t \ge c。 \end{cases} \tag{2}$$

The total time of feature vector B (N)=23P*P(O). Because the N P so when we choose namely the above namely the above, a feature vector of B K-means clustering algorithm, the total time running by (3n)O to P (3O), the efficiency of the algorithm is improved obviously. This low complexity algorithm is well suited for large data applications.

Spark is an efficient distributed computing system, compared to Hadoop, it is 100 times higher than the performance of Hadoop. Spark provides a higher level of API than Hadoop, the same algorithm implemented in Spark is often only Hadoop of the length of 1/10 or 1/100. Shark similar to the SQL on Spark is a data warehouse in Spark implementation, as is shown by equation(3), where f(t) is in the case of compatibility with Hive, the highest performance can reach one hundred times Hive.

$$f^1(t)=\sum_{i=i_1}^{i_2} b_i^1(t)k_i^1(t)x_i^1(t) \tag{3}$$

Big data market for new technologies, new products, new services, and new formats will continue to emerge. In the field of hardware and integrated equipment, big data will have an important impact on the chip, the storage industry, will also generate integrated data storage processing server, memory computing and other markets. In the field of software and services, big data will lead to the rapid processing of data analysis, data mining technology and the development of software products.

Big data not only refers to the massive data, also contains data segments, almost all of the links within the enterprise will be in the form of data to show, for example, efficiency optimization derived time node of the business aspects of the Amazon, in this area has been developed greatly, there will be a report and data processing operations based on a large number of every day operation strategy, marketing strategy, the change is mainly to see the data, automatic replenishment model it defined is the principle of time series and based on extreme value formed, effectively solve completely rely on the order, the artificial replenishment model to improve the efficiency of inventory management.

Electricity supplier is born with a big data aura. Compared to the traditional retail and channel, B2C platform can obtain consumer behavior data, shopping preferences, status, contact information, etc., for the user to accurately identify and locate [5]. On the platform of the third party service providers, such as logistics companies, payment companies, but also contributed to the operating conditions, product service records, consumer reviews, including important data. Do not have the advantage of big data, traditional businesses, on the one hand, self built electronic business platform, with the Internet and cooperation, to create their own customer management or membership system, on the other hand, with the opening of the O2O, such as the use of WIFI technology to create indoor interactive system, thus completing the data collection.

For large data in e-commerce applications can be an example, if the customer wants to buy a Chinese style doors and windows, he usually use search engines. When he entered the doors and windows, Chinese and other keywords, the search engine can display the relevant product information in the forefront of the search. For customers, large data makes him very precise access to the products they want, and Internet companies can analyze the user behavior, more accurately the needs of mining, with the use of large data analysis, a high degree of concern for the further promotion of investment products.

Individual performance, unstructured and semi-structured data, the general characteristic and basic principle is not clear, these are required by multi disciplines including mathematics, economics, sociology, computer science and management science, to study and discuss. Given a semi-structured or unstructured data, such as images, how to convert it into a multi - dimensional data table, object-oriented data model or directly based on the image data model, as is shown by equation(4).

$$w_{i+1}^1\left(t+1\right)=\left(1-wd_i^1\left(t\right)\right)x_i^1\left(t\right)-rs_i\alpha N^1\left(t\right) \tag{4}$$

The use of large data analysis to send personalized EDM. If customers have in the e-commerce website to view a product but did not buy, there are several possibilities: A. stock, B. price is not right, C. do not want to brand or not to goods, just to see if the D. in the customer view the goods out of stock is delivered immediately notify the customer; if there were goods and customers do not buy because the price is likely to be caused by, will inform the customer in the commodity price promotion; at the same time, inform the customer in the introduction and the similar goods or related commodity when warm [6]. In addition, through the excavation of the customer's periodic buying habits, in the vicinity of the customer's purchase cycle timely remind customers.

Shop operators need to analyze a lot of data, including external data, which is the industry's market share; it also reflects the website visits PV; reflect the electricity supplier website sales flow rate, order conversion rate, average order value and so on; of course the most important thing is lying on the Bank of the digital.

III. Novel Model of E-Commerce Marketing Based on Big Data Analysis and Processing

Electricity supplier big data collection, integration, analysis is a professional and rigorous method system, the traditional method is difficult to do, which requires the use of sophisticated computer software systems. Take Shanghai far Fang multi-user mall, its independent research and development of electricity supplier transaction data analysis module can be a very good statistical analysis of business data.

The emergence of the Internet and the development of related technologies make it possible to collect and analyze massive data. The characteristics of the Internet has led to the spread of these data can be high speed and large capacity [7]. The Internet introduces a pattern of data generated by the user. This model is characterized by multiple sources, low cost, and timelier. Of course, the authenticity and reliability of these data need to be certified. One of the advantages of building e-commerce based on the Internet and traditional retail is the availability of data. E-commerce can get the source of the customer's visit in real time, the search, collection, purchase behavior in the website, and the relevance of the goods purchased. These data can help companies more accurate customer service.

Application of electricity supplier data broadly divided into the following steps: A. data collection, verification and filtering; B. classification and stored in the data warehouse; C. data mining is to find association rules and data of the data between the D. data model and parameter establishment; adjustment; e. application development and decision support based on data.

The module consists of three parts: system statistics, sales data analysis, operational data analysis. Companies can easily access to the day of the site's traffic, website products, news and other published articles, online members, online visitors, etc.. As well as on-line members, and it is the current online membership calculation and the proportion of the total membership. At the same time, the module has the function of historical statistics, statistics of the past traffic data, with the trend chart in the form of display. The most important point is that the system can be presented for the mall, shop sales details, commodity classification statistical analysis, as is shown by equation (5) [8].

$$\frac{f'\left(x_1\right)}{2x_1}=(b^2+a^2)\frac{f'\left(x_2\right)}{4x_2^3}=\frac{\ln\frac{b}{a}}{b^2-a^2}x^3f'\left(x_3\right) \qquad (5)$$

Big data is a kind of artificial nature has the hidden law, searched for scientific mode of big data will bring a general method to study the beauty of big data, although this exploration is very difficult, but if we find the unstructured and semi-structured data conversion method of structured data, data mining methods known will become a major tool for data mining.

Big data is not in the computer CRM system and various forms, really big data is reflected in consumer behavior and feelings, all the bits exist in the form of Internet information, its basic features are: mass, high speed, flexibility, diversity! For example, you decide to buy is not the business of "pretty, but the evaluation of other buyers, the seller's credit rating, this is the big data; such as unknown problem, Baidu to search, today if there are a lot of people search for" how to treat colds, can predict ten days there will be a flu outbreak. This is called big data.

Using the data of the biggest weakness is the lack of relevance to grasp, once isolated data considerations, most core elements may be missing or not accurate and comprehensive expression of e-commerce internal information transfer can be transformed into data, by operating on the basis of the data association will become the basis for data analysis. Using the data of multi dimension and multi angle of view, through a core dimension of data will be gradually expanded, the reason will be a behavior and rationality through more than a dozen or more standard data to show, make it more accurate and focused.

The Internet is a National People's Congress in the age of media data, people through the computer and it together, not a historical product of large data is not because of strict environmental control and content, which fully reflects the dynamic data it is timely developed on the basis of the Internet content, big data information can be generated in any time is dynamic not only in the data collection process, data processing and data preservation technology are constantly changing, so it can be said that the facility is dynamic data processing.

According to statistics, 82% of the electricity supplier is being challenged to deal with massive amounts of information, and they spend a lot of time to study it, and 89% of the electricity supplier because of overload processing data and lost sales opportunities [9]. Just sitting on big data is not enough, the big data analysis and mining capabilities has become the core competitiveness of the electricity supplier. This shows that the key is not the number of large data in the number of raw materials, but the ability to process data, which can make the real value of big data.

Under the direction of big data products include: crowd, product features, and unique selling points (customized). Large number of products must be selected to do the STP analysis, the product must meet the following characteristics: 1) product profit space, for example, now sells stockings on Taobao, basically no profit margins. What do you do, three kinds of methods: one is to sell special products, for example, can sell sell sexy stockings at the speed, through the crowd of foreigners love China taste stockings. Two is engaged in wholesalers in Alibaba, Taobao market has been saturated, in doing this product is dead. The three is to increase product features, stockings with a special function of burning belly.

When the number of clusters is larger, that is, M is larger, the number of samples of the new training sample I NewX will be much less than the original sample number, it is easy to achieve a significant reduction in the amount of KNN calculation, improve the speed of classification purposes. However, with the increase of M, the overhead of clustering will increase, and the number of samples in the new training sample I NewX will gradually decrease, which may lead to the decrease of classification accuracy. So in order to avoid this situation, the number of clusters m need to be set in a more reasonable value, as shown in the following formula [10].

$$\mathrm{p}_1=\rho p_0, p_2=\frac{\rho^2}{2!}p_0,\cdots,p_n=\frac{\rho^n}{n!}p_0,p_{n+1}=\frac{\rho^{n+1}}{n\cdot n!}p_0,$$
$$p_{n+2}=\frac{\rho^{n+2}}{n^2\cdot n!}p_0,\cdots,p_{n+r}=\frac{\rho^{n+r}}{n^r\cdot n!}p_0,\cdots \qquad (6)$$

E-commerce marketing precision, as the electronic commerce platform profitable growth of consumer data, collected by the consumers from different channels of data, timely and accurate understanding of the comprehensive information of the customer. Especially the development of mobile smart devices can whenever and wherever possible to provide consumers with related services and products, provides a location data for the mobile users, on the other hand, because there are a large number of data in the intelligent mobile phone, is unique to the individual use, so that consumers one-on-one service possible.

In the case of similar classification time, KNN classification algorithm LSC based on clustering accuracy rate was significantly higher than that of KNN algorithm of random block, and can be found from table 8, the classification accuracy is closer to the classical KNN algorithm. Based on the above analysis, we can see that the number of clusters of the same size of M should be to ensure that all sub clusters can run in memory, as small as possible.

It can not only reduce the clustering time, but also improve the speed of classification, as is shown by equation (7).

$$I(t) = 1 - R_0 - S + \frac{1}{\sigma}\ln(\frac{S}{S_0}) \tag{7}$$

On the importance of enterprise data on the contrary, the massive data used by companies to do the absolute value ratio, add, subtract, multiply and divide method, trend, are the most frequently used methods, data is the abstraction and limitations of all a hideous mess, did not get a breakthrough; there are many reasons for this result, may be different stages of development, may is the lack of human resources, no matter what the reason is, a waste of such an important resource is a great loss to the enterprise, innovation in the field of data need to be improved.

Data audit includes two aspects, on the one hand is the integrity of the audit, as well as his right or wrong, or that he has not been unsafe to use. There may be a data produced, there will be an audit procedure to verify whether the data is correct or wrong, should not be produced at this time. E-commerce sites for big data security, but also the use of some conventional means. For example, it is the separation of internal and external networks and it is a data management process is being set up, including large data audits. When we have just discovered that a data has been tampered with, or illegally used, the audit mechanism will start the alarm.

In the development of big data technology today, more and more e-commerce companies have to use big data to seek their own business development and talent needs. Many colleges and universities are based on the creation of e-commerce curriculum corresponding social development, but in the professional training program design is often biased in favor of the popular professional, lack of electronic business enterprise and the social demand for talent data collection, analysis. The use of big data technology will be able to better provide a more accurate and reliable data basis for personnel training programs. The main data collection requirements, social classification and acquisition of electronic commerce environment in general post supply demand; electronic business enterprise of professional talents, skills and quality requirements; school teachers professional settings, and other data; student information, interest, skill biased data.

IV. SUMMARY

Big data has changed the traditional mode of management decision-making structure. The study of the impact of big data on management decision making will be an open research question. In addition, the change of the decision-making structure requires people to explore how to support the higher level of decision-making and do the two mining". No matter what kind of data heterogeneity big data, big data in the rough knowledge can still be seen as a mining category. It is very necessary to find the bridge between the heterogeneity of data and the heterogeneity of decision making by looking for the "knowledge" generated by the two mining. How to change the decision structure under the condition of big data is equivalent to the study of how to make the decision maker's subjective knowledge in the process of decision making.

The paper presents novel model of e-commerce marketing based on big data analysis and processing. E-commerce marketing, has been changing with the spread of the way, from the traditional to the network, and then from the network to social, and then from PC to mobile, each change is profound. Every technological progress also brings the progress of e-commerce industry! The main characteristics of e-commerce network marketing in the era of big data is accurate demand forecasting, mining, interpersonal orientation, directional communication, and ultimately by the user to grasp the decision-making power and the two transmission right, rather than through its own e-commerce business. Therefore, e-commerce network will usher in the era of big data interaction.

ACKNOWLEDGMENT

This paper is supported by the National Natural Science Funds of China (61272015), and also is supported by the science and technology research major project of Henan province Education Department (13B520155, 17B520026) and Henan Province basic and frontier technology research project (142300410303).

REFERENCES

[1] Sirivimol Thanchalatudom, Namfon Assawamekin, Using Big Data Technology for Information Management in Hybrid Learning System, *RNIS*, Vol. 12, pp. 179 -182, 2013.

[2] Hyoung woo Park, Il Yeon Yeo, Haengjin Jang, Seo-Young Noh, “Simulation based Analysis on Big Data Service Bottleneck for Data Center”, JNIT, Vol. 4, No. 8, (2013),pp. 185 – 189.

[3] Hongxin Wan, Yun Peng, "Clustering and Evaluation on Electronic Commerce Customers Based on Fuzzy Set", IJACT, Vol. 5, No. 3, pp. 199 ~ 206, 2013.

[4] Wei Li, Hongtu Zhang, Tingting An, "Optimal Decision-Making in E-Commerce Platform Based On Optimal Stopping Theory", JCIT, Vol. 8, No. 8, pp. 922 ~ 929, 2013.

[5] Dawei Sun, Ge Fu, Xinran Liu, Hong Zhang, "Optimizing Data Stream Graph for Big Data Stream Computing in Cloud Datacenter Environments", IJACT, Vol. 6, No. 5, pp. 53 ~ 65, 2014.

[6] Shoupu Wan, Yongjia Wang, Michael Recce, “A Case Study of the Internet Dynamics -From Big Data to Marketing Insights”, IJIIP, Vol. 5, No. 2, (2014), pp. 39-52.

[7] JinHui Lei, XianFeng Yang, "Construction the E-commerce Trading Platform Based on Rough Set Data Mining Technology", JCIT, Vol. 8, No. 3, pp. 460 ~ 469, 2013.

[8] Sun Chengshuang, Shen Liyin, Gan Lin, "An Analysis on Competitiveness Enhancement by Applying E-commerce to Construction Corporations", JDCTA, Vol. 7, No. 8, pp. 772 ~ 780, 2013.

[9] Fan Yang, "The Research of E-Commerce Recommendation System Based on Cloud Computing", IJACT, Vol. 4, No. 16, pp. 256 ~ 263, 2012.

[10] Yanqing Lv, Jianmin Gao, Zhiyong Gao and Hongquan Jiang, “Multifractal information fusion based condition diagnosis for process complex”, Process Mechanical Engineering, pp.1-8. (2012),

Research on Tool Path Planning Method of NURBS Surface Based on CPU - GPU Parallel Computing

Wujia Yu
School of Automation, Hangzhou University of
Electronic Science and Technology HDU
Hangzhou, China
E-mail: yuwujia@163.com

Zhendong Li
School of Automation, Hangzhou University of
Electronic Science and Technology HDU
Hangzhou, China
E-mail: 57815697@163.com

Yangqiang Bi*
School of Automation, Hangzhou University of
Electronic Science and Technology HDU
Hangzhou, China
E-mail: byq_work@163.com
*The corresponding author

***Abstract*—In order to deal with the inefficiency of trational serial tool path algorithms and incompatibility issues on the heterogeneous hardware platforms, this paper suggests a tool path planning method based on CPU-GPU(Central Processing Unit-Graphic Processing Unit) heterogeneous parallel computing. The method contra poses NURBS(Non-Uniform Rational B-Splines) surface which is abstracted as a matrix multiplication on the principle of isoparametric line tool path planning method. Then a parallel algorithm in accordance with Open CL(Open Computing Language) specification is proposed. Adopting data parallel programming model, the method executes multiple work-items of the GPU on the core under control of the CPU logic, and reconstructs the isoparametric line method as parallel execution instead of traditional serial execution. Simulation results show that this algorithm takes less time to generate tool paths on the CPU-GPU heterogeneous platforms, reduced by 1.5 to 15.9 times compared with traditional serial algorithm and it is of great significance to the tool path planning's real-time or quasi real-time generation.**

Keywords-Component; Nurbs surfaces; OpenCL; Parallel computing; Tool path planning; CPU-GPU

I. INTRODUCTION

In recent years, the parallelization of CNC(Computer numerical control) system computing task has been developed rapidly. The traditional CNC software architecture and computing model did not consider about the problem of parallelization at the beginning of the design, especially in the tool path planning, the traditional methods are usually used in the design of serial computing[1-2]. And the introduction of parallel computing technology can enhance the performance of the tool path planning algorithm compared with the traditional serial computing method. [3] studied the problem of parallel processing CNC system core mandate and the establishment of a parallel evaluation model from a system perspective. [4] proposed the tool path generation algorithm based on parallel computing, and discussed the parallelization problem of tool path planning algorithm.

As the above is usually the software-level of parallel technology, real parallel computing is a hardware-level multi-threaded parallel computing and heterogeneous systems of parallel computing, it has never been involved in the field of NC[5]. In this paper, the use of OpenCL to the CPU and GPU which are different processor architectures, can be calculated synergistically, the CPU for logic control and serial computing, the GPU multiple work-items perform the same kernel program to solve the NURBS surface tool path, building up the parallel computing of heterogeneous system. OpenCL is an open, parallel computing standard for cross-platform, play a role in the different performance of the new hardware, so that the parallel algorithm has better openness and compatibility[6-7].

II. BASIC PRINCIPLE

A. Tool Path Planning Method

Tool path planning method can be divided into three categories: cross-section method, projection method, parametric line method[8]. This paper only using the parametric line method to carry on the parallelization research. The parametric line method uses one parameter direction of the surface as the direction of tool, the other parameter direction is the feed direction, and then the processing surface along the selected direction in the parameter domain within the parameters of subdivision, the formation of a number of parametric line information, and finally in accordance with certain rules connected parametric line nodes constitute the tool path.

For NURBS surface, which STEP(Standard for the Exchange of Product Model Data) defines it as the only mathematical method for industrial product geometry, it is assumed that it is a mesh surface consisting of a series of NURBS curves with constant u value and v value[9].

Therefore, the solution of NURBS surfaces can be decomposed into the grid point composed of NURBS curve corresponding to each *u* value and *v* value of node vector, which is consistent with isoparametric line tool path planning method. Therefore, for NURBS surface, the tool can select the direction along the surface (*u* or *v*) as the direction of tool, (*v* or *u*) direction as the direction of feed and tool path is generated by the isoparametric method. After subdividing the parameters *u*, *v*, we can get a series of *u* or *v* parameter line. Since each of the *u*-parameter line or the *v*-parameter line is not related, each of the *u*-parameter line or the *v*-parameter line corresponds to a series of coordinate values of the *v*-parameter or *u*-parameter subdivision, that is coordinate information of the grid points can be calculated by parallel calculation. In this paper, the *u* direction is taken as the feed direction, and the *v* direction is used as the tool direction and the tool path planning is carried out by the isoparametric method.

B. *CPU-GPU Heterogeneous System*

In the CPU-GPU heterogeneous system, CPU is a multi-instruction single-data stream architecture, the data processing is basically a single pipeline, which is very good at doing logical control, while the GPU is a typical single-instruction multiple-data architecture, which is specialized in data calculation[10]. Heterogeneous system architecture shown in Fig .1, CPU and GPU through the external bus interconnection, each with its own storage space, respectively, the main memory and video memory. The execution of the program can be roughly divided into three steps in the CPU-GPU heterogeneous system: first, the input data from the CPU main memory copy to the GPU video memory; then, call the GPU implementation; Finally, the calculation results from the GPU video memory copy back to the CPU main memory.

OpenCL as a new parallel computing technology, it can call all the computer computing resources, including CPU, GPU and other processors. In the execution model of OpenCL, the program is divided into two parts to execute, namely the main program and the kernel program, in which the main program runs on the host computer, the kernel program runs on the OpenCL device (computing device), and the main program manages the running of the kernel program.

III. DESIGN AND IMPLEMENTATION OF PARALLEL POOL PATH PLANNING ALGORITHM

A. *Algorithm Design*

The tool path of the NURBS surface is correct cutting tools, which is composed of a series of grid points of the coordinate information in accordance with the provisions of the tool direction and feed direction. The row vector **U**, which consists of all the u-subdivided values, are combined with a large matrix, and the column vector $\mathbf{V}^{\mathbf{T}}$ consists of all *v*-subdivisions are also represented by a large matrix, where each subdivision value corresponds to a parametric line, furthermore, the coefficient matrix is **M**, the control vertices are denoted by the matrix **G**, w_i denoted by the matrix **W**, so the NURBS surface equation can be regarded as a series of matrix operations, the denominator is a series of matrix multiplication, similarly, the numerator is also a number of columns matrix multiplication and then inverse operation, and finally multiplied by the denominator to get the final grid point coordinate information. As the parameters *u* and *v* is not related, and is independent, they can use the assigned computing unit and the processing unit. Each processing unit performs the same matrix multiplication kernel function in parallel to get the surface parameter grid point information.

B. *The OpenCL Implementation of the Algorithm*

According to the algorithm, the computational complexity of the coefficient matrix multiplied by the control vertex matrix is small, arranged on the CPU. However, the multiplication of the coefficient matrix and the parameter matrix is computationally large and can be calculated in parallel, so it is executed on the GPU.

The main program transforms the elements which are composed of the subdivision values *u*, *v* in the parameter matrix **U**, **V** into one-dimensional arrays U[i], V[i], storing them in the CPU memory, each subdivision *u*, *v* in the parameter matrix corresponds to a parametric line; The U[i], V[i] array in the CPU memory is then copied into the OpenCL memory object buffer of the applied memory space, which corresponds to the buffer context and the context associated with the device (GPU) is globally visible, and the buffer is set to a parallel access to the memory, which can reduce the access between different threads conflict or blocking, thereby reducing the data transmission and communication overhead; And the main program obtain data from the buffer to the kernel and then executes it on the CPU. The implementation of multiple work-items in parallel with the same kernel, the result is the information of all the parametric line, which the result calculated for each work-item corresponds to a row element in the matrix, that is the coordinate information of a parametric line; The final result is also written to the global result buffer, and then the data of the result buffer through a specific function mapped to the host memory available to other parts of the program.

The steps of using OpenCL to design the main program and kernel[11]:

1) Write the kernel function KERNEL(_kernel voidMatrixmultipli()) according to the matrix multiplication logic;

2) Call clGetPlatformIDs() to find OpenCL platform collection in the computer system;

3) Through the clCreateContextFromType() to establish context for the communication between the computing device , the memory object and the command queue;

4) clCreateProgramWithSource() creates a program object associated with context and use clBuildProgram() to build program object for the specified device;

5) Create CreateKernel() on the device, execute the kernel in parallel, the kernel is __kernel void Matrixmultipli ();

6) clCreateCommandQueue() to create a command queue, submit a command to the command queue to complete the specific operation of OpenCL;

7) Call clCreateBuffer() to create a buffer to facilitate the data read and write, through the clSetKernelArg() passes the kernel parameters and buffers to the kernel;

8) Call clEnqueueWriteBuffer() the data to be involved in the calculation of information written to the buffer, here is u, v subdivided one-dimensional array;

9) clEnqueueNDRangeKernel() added the kernel event to the command queue, ready to perform on the GPU, the function parameters set the size of the workgroup and work-items;

10) Call the clEnqueueMapBuffer() and memcpy() map the buffer data to the host;

Follow the steps above to design the program and execute it on the GPU. The CPU gets the coordinate information of all the parameter grid points.

IV. Experimental Results and Analysis

The speedup is one class of the criteria for measuring the performance and effectiveness of program parallelism. It is usually defined as the ratio of the serial program execution time to the parallel program execution time:

$$\text{Speedup Ratio} = \text{Serial Program Execution Time} / \text{Parallel Program Execution Time} \qquad (1)$$

The test environment of this paper are the AMD(Advanced Micro Devices) Radeon (TM) R9 200 series GPU, video memory 2048MB, RAM(Random Access Memory) 2G, CPU using AMD Athlon (TM) 3.7GHz. In order to simulate the performance of the program in different computational scale, NUBRS surface is sampled at different sampling densities, that is through different *u*, *v* subdivision, statistics on the tool path generation time. In order to reduce the interference with other programs in the system to the operation of the algorithm, the expression (1) takes the method of repeat the three operations to take the arithmetic squared value as the experimental value, the numerical result retains one decimal place. The experimental results are shown in Table I.

From the experimental results, we can see that the execution efficiency of parallel program under heterogeneous platform is improved compared with traditional serial algorithm. Through the speedup ratio and the implementation of time we can conclude that when the number of points are small, the acceleration effect is not very obvious; when the points continue to increase, the parallel computing gradually reflected the advantages.

After the *u*, *v* subdivision is greater than 1000, the tool path is too dense and the effect is not obvious. So only the parameters *u*, *v* fineness 40 * 40,100 * 100,1000 * 1000 uniform sampling tool path simulation diagram. As the picture shows:

Through simulation that when the parameters are 40 * 40, the serial execution time of 0.625ms, parallel execution time of 0.352ms, the effect of the general. When the parameter is 100 * 100, the serial execution time is 2.241ms, the parallel execution time is 0.864ms, the effect is obvious. When the parameter is 1000 * 1000, the serial execution time is 198.517ms, the parallel execution time is 53.781ms, the effect is remarkable.

V. Conclusion

Based on NURBS surface characteristics, this paper uses OpenCL technology to establish a tool path planning method based on CPU-GPU heterogeneous parallel computing. Through the analysis of the experimental results, the parallelization of the calculation greatly enhance the performance of the isoparametric line planning method, and realize the cooperative parallel computing of different architectures.

TABLE I. NURBS Surface Two Kinds of Program Execution Time (Unit: ms)

U fine fraction * V fine fraction	Serial program execution time	Parallel program execution time	Speedup ratio
40*40	0.6	0.4	1.5
100*100	2.2	0.9	2.4
1000*1000	198.6	53.8	3.7
2000*2000	768.4	96.7	7.9
3000*3000	1768.5	138.6	12.8
4000*4000	3026.8	198.4	15.2
5000*5000	4897.1	306.3	15.9

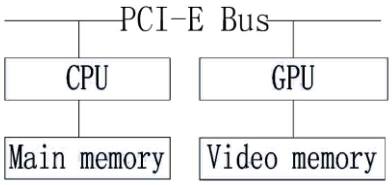


Figure 1. CPU-GPU heterogeneous architecture diagram

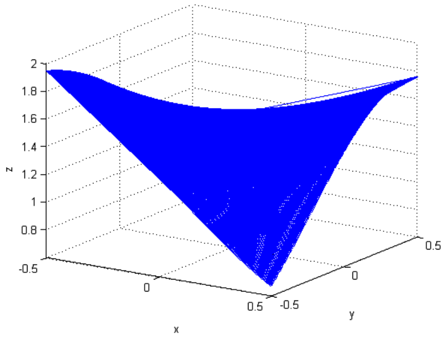


Figure 2. NURBS surface example.

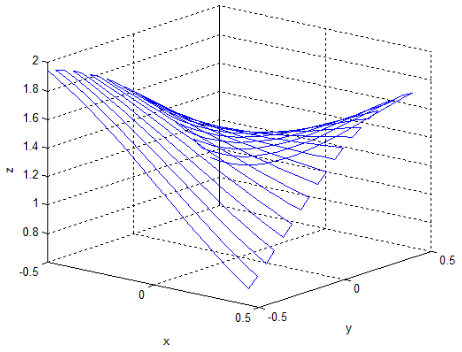


Figure 3. 40* 40 uniform sampling.

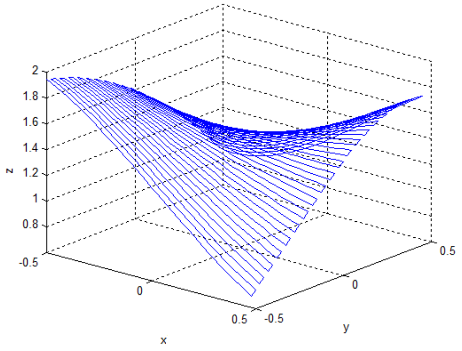


Figure 4. 100*100 uniform sampling.

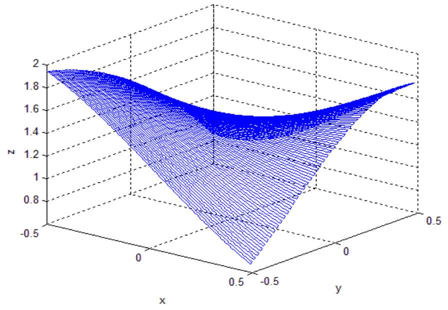


Figure 5. 1000*1000 uniform sampling.

Acknowledgment

This work is partially supported by the National Natural Science Foundation of China (5140051146).

References

[1] Qin Hong-xin, Hua Rui. Research on Tool Path Planning for NC Machining of Freeform Surface[J]. Coal Technology 2013,30(3):40-41

[2] Li li, Fang li-jin, Wang Guo-xun. Research on Tool Path Planning of NURBS Surface Five - axis Machining[J]. Machinery 2014,52(2): 5-9.

[3] Zhang Xiang-li, Tang Xiao-qi, Chen Ji-hong. Parallel processing of computer numerical control system [J] .Computer Integrated Manufacturing System, 2008,14(8), 1603-1607.

[4] Yu Zhan-yue, Zhou Ruo-rong, Zhuang Hai-jun. A Parallel Algorithm for Tool Path Generation in NC Machining [J]. Mechanical Science and Technology, 2004,23 (3), 266-268.

[5] Yu Wu-jia. STEP-NC-based five-axis machining tool path planning method [D]. Doctoral dissertation, Zhejiang University.

[6] Stone J E, Gohara D, Shi G. OpenCL: A parallel programming standard for heterogeneous computing systems[J]. Computing in science & engineering, 2010, 12(3): 66-73.

[7] Du P, Weber R, Luszczek P, et al. From CUDA to OpenCL: Towards a performance-portable solution for multi-platform GPU programming[J]. Parallel Computing, 2012, 38(8): 391-407

[8] WU Fu-zhong.Modeling Path Planning of Equal Spacing Tool for Free-form Surface Coordinate Processing [J] .Computer Integrated Manufacturing System, 2007, (10): 2064-2070.

[9] Starly B, Lau A, Sun W, et al. Direct slicing of STEP based NURBS models for layered manufacturing[J]. Computer-Aided Design, 2005, 37(4): 387-397.

[10] Research on Virtualization Technology Based on CPU / GPU Platform [D].Journal of Shanghai Jiaotong University

[11] Munshi A, Gaster B, Mattson T G, et al. OpenCL programming guide[M]. Pearson Education, 2011:25-73.

Discussion on the Factors of Stability for Benchmark Example with a Spherical Failure Surface in Clay

Jinhui Liu
College of traffic and civil engineering
Shandong Jiaotong University
Jinan, China
Liumengdi1996@163.com

Wantao Ding*
Geotechnical & Structural Engineering Research Center
Shandong University
Jinan, China
dingwantao@sdu.edu.cn
*The corresponding author

***Abstract*—A three-dimensional slope stability problem involving a spherical failure surface in clay is often used in the literature as a benchmark example against which numerical models are validated. In the existing research literature, the analytical expression has been obtained for the factors of safety by assuming plane-strain mechanisms during slope failure. And the hypothesis does not comply with the actual project due to the size effect of slope and surrounding constraints placed on the slope. This paper compares and analyzes the results of the existing research literature. And the three analytical expressions under the three kinds of rotational model have been given for the factors of safety. In practice, the value of factor of stability obtained by numerical model should be in a range which is determined by rotational model of failure body. In this paper, when subtended angel is equal to 60°, the value of factor of stability obtained by numerical model for benchmark example should be reasonable in the range from 1.191 to 1.588.**

Keywords-Three-dimensional slope stability analysis; Factor of safety; Spherical failure surface; Numerical model; Plane-strain mechanisms; Subtended angle

I. INTRODUCTION

Because of occupying a certain space, failure surface of slope has obvious three-dimensional feature. However, its stability analyses have usually been carried out using a two-dimensional approach. Such failure modes yield conservative estimates of the slope safety when compared with three-dimensional failure patterns. Since the mid 1970s, increasing attention has been directed toward the implementation of three-dimensional stability models [1,2]. Three-dimensional analyses of slopes can be grouped into three categories: the extension of traditional slice methods; numerical approaches, such as the finite element method or the discrete element mechod; and limit analysis (the plasticity approach). The reader will find a review of the first two categories in a recent article by Griffiths & Marquez [3]. The application of limit analysis to earth slopes started with a paper by Drucker & Prager [4], who applied the kinematic approach of limit analysis to the stability of slopes undergoing plane-strain failure. Where the limit equilibrium methods of columns are most popular and are considered to be the most feasible for practical problems [5-9]. L.Z. Wu. et al. conducted a series of physical tests, which were conducted to simulate rain-induced slope failure[10]. Jin Man Kim proposed the reliability approach to analyze slope stability with spatially correlated soil properties[11]. A. Johari used the jointly distributed random variables method to assess the reliability of infinite slope stability[12]. Seboong Oh present two case studies of rainfall-induced failure of engineered slopes[13]. Joshua A. White assessed slope stability using stochastic rainfall simulation[14]. L.L.Zhang reviewed the stability analysis of rainfall-induced slope failure[15]. In reality, under rainfall conditions, the degree of saturation within a slope and along a failure surface could be highly variable and pore water pressure could be negative[16]. In recent years, slope-stability analyses have been expanded to include coupled hydromechanical processes under variably saturated conditions[17-18].

Due to the complexity of soil, the analytical solution of some actual problems can’t be obtained. Numerical methods known as a viable option need to be validated using a comparison with a particular problem that involves a spherical failure surface in clay for which a closed-form solution appeared to be available. However, after reconsideration of this benchmark example, it was found that the reported value of the factor of safety obtained from the closed-form solution was several problems. One is the issue of the coordinate system used; the other is the issue of the rotational problems of failure body. By analyzing the existing research methods and considering the effect of slope spatial dimensions during the slope failure, this paper discusses the potential sliding type of failure surface in clay and obtains the factors of safety under different sliding type.

II. ANALYSIS

A. Arm of Resisting Moment

When calculating the resisting moment M_{ρ}^{o}, Silvestri [19] considers the moment arm about the axis OA (z-axis, Fig .1) is defined as

$$L_1 = R\cos\psi \tag{1}$$

In which R is the spherical radius, ψ is the angle between the radius R and the y-axis in the yz-plane. In fact, the moment arm L_1 does not rotate about the z-axis, as shown in Fig .4. It represents the moment arm which rotates the axis (connection line from B to C,eg. BC) parallel to the z-axis. The moment arm L is given by

$$L = R\sqrt{\sin^2\theta + \cos^2\theta\cos^2\psi} \tag{2}$$

Also, eq.(2) can be written as

$$L = R\sqrt{1-\cos^2\theta\sin^2\psi} \tag{3}$$

Where θ is the angle between the radius R and the y-axis in the xy-plane.

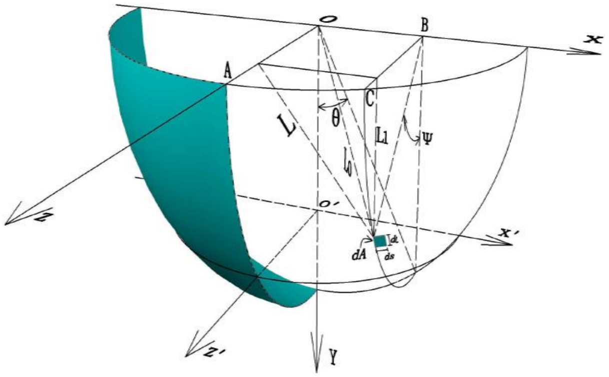


Figure 1. Moment arm of infinitesimal area element under spherical coordinates

B. Factor of Safety

The mechanism of failure consists of a rigid body rotation about the axis OA (z-axis, Fig .2a). The driving moment M_d^o is defined as

$$M_d^o = \gamma\frac{\pi R^4}{4}\cos^4\delta\sin\beta \tag{4}$$

Where γ is the unit weight, β is the inclination angle of the slope, and δ is shown in Fig .2(a).

dA is the infinitesimal element of spherical surface involved in the slide, as shown in Fig .2(b) and Fig .4. dA can be expressed as

$$dA = dlds \tag{5}$$

Where $dl = R\cos\theta d\alpha$ and $ds = Rd\theta$, as shown in Fig .2(b). Or $dl = R\cos\theta d\psi$ and $ds = Rd\theta$, as shown in Fig .4, then eq.(5) reduces to

$$dA = R^2\cos\theta d\alpha d\theta \tag{6}$$

The moment arm rotation about the z-axis of the infinitesimal element dA is defined as

$$L_1 = R\sqrt{1-\sin^2\alpha\cos^2\theta} \tag{7}$$

And the resisting moment M_r^o is given by the following relationship

$$M_r^0 = 4S_u R^3\int_\delta^{\pi/2}\int_0^{\pi/2}\cos\theta\sqrt{1-\sin^2\alpha\cos^2\theta}\,d\alpha d\theta \tag{8}$$

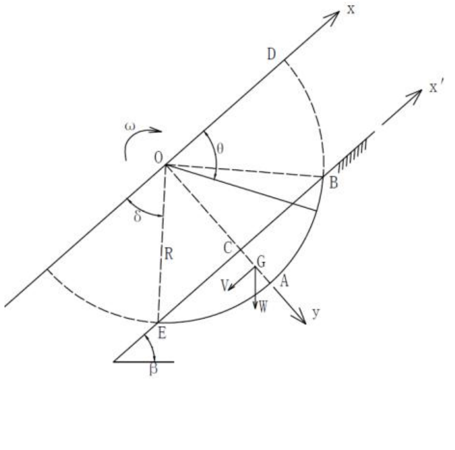
As a consequence, the factor of safety F is

$$F = \frac{M_r^o}{M_d^o} = \frac{16S_u}{\gamma\pi R\cos^4\delta\sin\beta}A(\theta,\alpha)$$

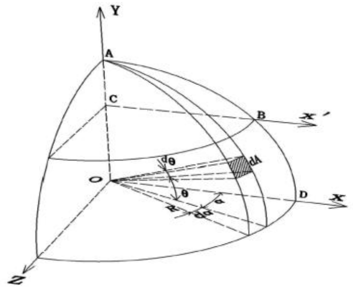
□ □ □ □ □ □ □ □
□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

$$F = \frac{M_r^o}{M_d^o} = \frac{16S_u}{\gamma\pi R\cos^4\delta\sin\beta}A(\theta,\alpha) \tag{9}$$

Where: $A(\theta,\alpha) = \int\limits_{\theta=\delta}^{\theta=\pi/2}\int\limits_{\alpha=0}^{\alpha=\pi/2}\cos\theta\sqrt{1-\sin^2\alpha\cos^2\theta}\,d\alpha d\theta$



(a) Cross-section of mechanism



(b) Coordinate system

Figure 2. Spherical cap failure surface: (a) cross-section of mechanism; (b) coordinate systemDiscussion

Because of occupying a certain space, the failure evolution process of slope has the obvious spatial characteristics. When assuming that the size of purely cohesive slope tends to infinity along the longitudinal and transverse section, the failure mode of slope can be considered plane-strain mechanisms. However, the plane-strain mechanisms of failure is more difficult to be met in the actual engineering, the width and length of the actual slope are limited. So the failure model of slope does not satisfy the plane-strain mechanisms. In the past, considering the plane-strain mechanisms, the hypothesis that the mechanism of failure consists of a rigid body rotation about the axis OA (z-axis, Fig.1) is usually proposed when analyzing the slope stability. Considering the size effect of slope and surrounding constraints placed on the slope, there may be a variety of rotational forms. Analyzing the evolution of slope slip, there are the other two kinds of rotational form. One is that all of the infinitesimal area elements rotate about point O (eg. Moment arm L_0 in Fig.1) ; the other is that all of the infinitesimal area elements rotate about the axis parallel to z-axis(eg. moment arm L_1 in Fig.1).

C. *All of the Infinitesimal Area Elements Rotation about Point O*

In this rotational form, as shown in Fig .1, the infinitesimal area element dA can be express as eq.(5). Where $dl = R\cos\theta d\psi$ and $ds = Rd\theta$, then eq.(5) reduces to

$$dA = R^2 \cos\theta d\psi d\theta \tag{10}$$

The moment arm rotation about point O is given by

$$L_0 = R \tag{11}$$

And the resisting moment M_r^o is given by the following relationship

$$M_r^0 = 4S_u R^3 \frac{\pi}{2}(1-\sin\delta) \tag{12}$$

As a consequence, the factor of safety F is

$$F = \frac{M_r^o}{M_d^o} = \frac{16S_u}{\gamma R\pi\sin\beta}\frac{\pi(1-\sin\delta)}{2\cos^4\delta} \tag{13}$$

D. *All of the Infinitesimal Area Elements Rotation about the Axis Parallel to Z-axis*

In this rotational form, the infinitesimal area element dA can be express as eq.(10).The moment arm rotation about the axis parallel to z-axis is given by

$$L_1 = R\sin\theta \tag{14}$$

And the resisting moment M_r^o is given by the following relationship

$$M_r^0 = 4S_u R^3 \frac{\pi(1-\sin^2\delta)}{4} \tag{15}$$

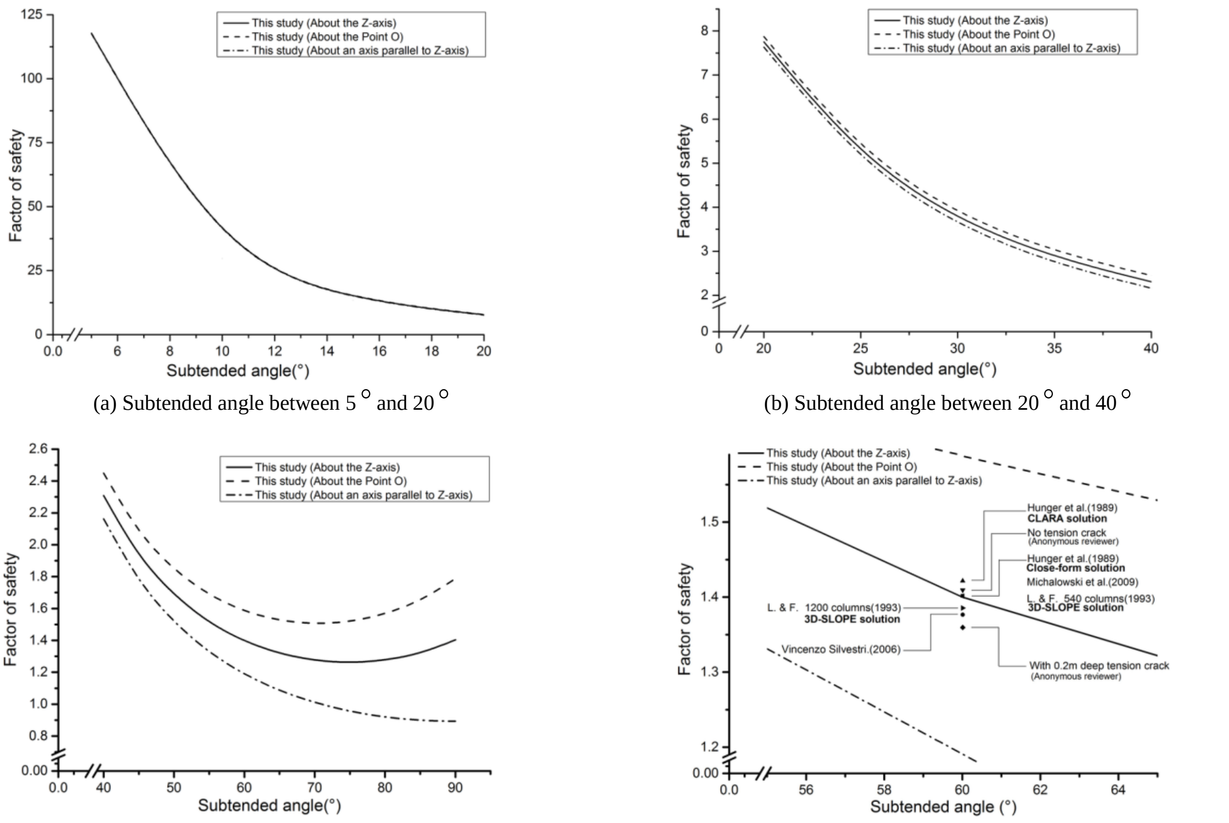
As a consequence, the factor of safety F is

$$F = \frac{M_r^o}{M_d^o} = \frac{16S_u}{\pi\gamma R\sin\beta}\frac{\pi}{4\cos^2\delta} \tag{16}$$

III. Comparison and Application

The geometry and parameters of the benchmark example used in the evaluation of the numerical approaches are (Hungr et al.1989; Lam and Fredlund 1993): $\beta = 26.6°$, $\Theta = 60° = \pi/3$, $S_u/\gamma R = 0.1$, $\delta = \pi/2 - \Theta$. In which Θ represents the subtended angle, as shown in Fig .1(a).

The range of subtended angle Θ is between 0 ° and 90 ° . Changing the value of subtended angle at intervals of 5 degree, the factors of safety rotation about axis can be obtained using the eq.(9), eq.(13) and eq.(16). The relations between the factors of safety and subtended angles are shown in Fig .3.



(a) Subtended angle between 5 ° and 20 °

(b) Subtended angle between 20 ° and 40 °

(c) Subtended angle between 40 ° and 90 °

(d) Comparison of factors of safety under subtended angel equal to 60 °

Figure 3. Factor of safety as function of subtended angle for benchmark example

As shown in Fig .3(a) and Fig .3(b), the factors of safety decrease gradually with subtended angels increasing. And the factor of safety rotation about point O is the largest. That rotation about axis parallel to z-axis is the least and that rotation about z-axis is the middle. As shown in Fig .3(c), the factor of safety rotation about axis parallel to z-axis decrease gradually with subtended angles increasing, but the other two factors of safety decrease gradually with subtended angles increasing from 40 ° to about 75 ° , when the subtended angles large than 75 ° , both of the factors of safety increase gradually.

As shown in Fig .3(d), Hungr et al. mentioned that while the closed-form solution of Baligh and Azzouz yielded F =1.402, the numerical model CLARA resulted in F =1.422. Considering a kinematically admissible rotational mechanism in cohesive soils (undrained behavior), Michalowski et al. yield F =1.402. One of the two anonymous reviewers of the present paper obtained F =1.41 using Bishop’s simplified method, as implemented in the latest version of the program CLARA. However, the program indicated negative normal stresses in 15% of the weight. With a vertical, planar dry tension crack 0.2m deep, the factor of safety reduced to 1.36, with the negative stresses affecting only 9% of the slide. In addition, Lam and Fredlund, using the 3D-SLOPE model, obtained F =1.402 when the slope was discretized 540 columns and F =1.386 for 1200 columns. The latter result puzzled Lam and Fredlund because it was lower than the so-called exact value of 1.402 when the number of columns was increased. However, in comparison with the value of F =1.191 obtained in this study(rotation about an axis parallel to z-axis), the factor of safety proposed by Lam and Fredlund, that is, F =1.386, is reasonable. Silvestri obtained F =1.377, the value is in inaccurate. In this study, three factors of safety were proposed: one is F =1.4(rotation about z-axis), the other is F =1.588(rotation about point O) and the third is F =1.191(rotation about axis parallel to z-axis). In past, the so-called exact value of F was considered equal to 1.402 because of assuming plane-strain mechanisms during slope failure. In the actual project, the slope failure does not satisfy plane-strain mechanisms completely, so the value of factor of stability for benchmark example using numerical models should be reasonable in the range from 1.191 to 1.588.

IV. Conclusion

A benchmark example often used to validate numerical models for the analysis of three-dimensional slope stability problems, was re-analyzed, and the analytical expressions under three kinds of rotational models have been obtained for the factors of safety. Assuming plane-strain mechanisms during slope failure, the factor of safety should be determined by eq.(9). But in the actual project, because of the size effect of slope and surrounding constraints placed on the slope, the three-dimensional numerical approaches yield solutions should be in a range between the value of factor of safety obtained by eq.(16) and that obtained by eq.(13).

Acknowledgment

The authors acknowledge the financial support provided by the National Natural Science Foundation of China (Grant Nos. 41572275) and the Shandong Province Natural Science Fund (Grant Nos. ZR2012EEM006).

Reference

[1] Baligh, M. M. & Azzouz, A. S.. End effects on stability of cohesive slopes. ASCE J. Geotech. Engng Div.101, No.GT11, 1975, pp.1105-1117.

[2] Hovland, H.J. Three-dimensional slope stability analysis method. ASCE Journal of the Geotechnical Engineering Divisioon, 103(9),1977,pp.971-986.

[3] Griffiths, D.V. & Marquez, R.M. Three-dimensional slope stability analysis by elasto-plastic finite elements. Géotechnique 57, No.6,2007,pp. 537-546.

[4] Chen, R.H., and Chameau, J.L.Three-dimensional limit equilibrium analysis of slopes. Géotechnique,32(1),1982,pp.31-40.

[5] Cavounidis,S. On the ratio of factor of safety in slope stability analyses. Géotechnique, 37(1),1987,pp.207-210.

[6] Hungr, O. An extension of Bishop's simplified method of slope stability analysis to three dimensions. Géotechnique, 37(1), 1987, pp.113-117.

[7] Gens,A., Hutchison, J.N., and Cavounidis, S. Three-dimensional analysis of slopes in cohesive soils. Géotechnique, 38(1), 1988, pp.1-23.

[8] Hungr, O., Salgado, F.M., and Byrne, P.M. Evaluation of a three-dimensional method of slope stability analysis. Candian Geotechnical Journal, 26(4),1989,pp.679-686.

[9] Michalowski, R.L., and Drescher,A. Three-dimensional stability of slopes and excavations. Géotechnique, 59(10), 2009, pp.839-850.

[10] L.Z.Wu, R.Q.Huang,Q. Xu,et al. Analysis of physical testing of rainfall-induced soil slope failures. Environ Earth Sci 73,2015,pp.8519-8513.

[11] Jin Man Kim, Nicholas Sitar. Reliability approach to slope stability analysis with spatially correlated soil properties. Soils and Foundations,53(1),2013,pp.1-10.

[12] A Johari, A.A. Javadi. Reliability assessment of infinite slope stability using the jointly distributed random variables method. Scientia Iranica A, 19(3),2012,pp.423-429.

[13] Seboong Oh, Ning Lu.Slope stability analysis under unsaturated conditions: Case studies of rainfall-induced failure of cut slopes. Engineering Geology, 184,2015,pp.96-103.

[14] Joshua A. White, Dashi I. Singham. Slope stability assessment using stochastic rainfall simulation. International Conference on Computational Science, ICCS 2012,pp.699-706.

[15] L.L.Zhang, J.Zhang, L.M. Zhang et al. Stability analysis of rainfall-induced slope failure:a review. Geotechnical Engineering, Issue GE5,V(164),2011,pp.299-316.

[16] Buscarnera,G., Whittle, A. Constitutive modelling approach for evaluating the triggering of flow slides. Can. Geotech. J.49(5),2012,pp.499-511.

[17] Lu, N., Kaya, M. A drying cake mechod for measuring suction stress characteristic curve, soil-water retentaion, and hydraulic conductivity function. Geotech. Test.J. 36,2012,pp.1-19.

[18] Lu, N., Wayllace, A., Oh, S. infiltration-inducd seasonally reactivated instability of a highway embankment near the Eisenhower Tunne, Colorado, USA. Eng. Geol.162,2013,pp.22-32.

[19] Vincenzo Silvestri. A three-dimensional slope stability problem in clay.Can. Geotech. J. 43,2006,pp.224-228.

Intrusion Detection Based on Self-adaptive Differential Evolutionary Extreme Learning Machine

Junhua Ku
Department of information engineering
Hainan institute of science and technology
Haikou, China
E-mail: kujunhua@163.com

Dawei Yun
Department of information engineering
Hainan institute of science and technology
Haikou, China
E-mail: cogemm@163.com

Bing Zheng
Department of information engineering
Hainan institute of science and technology
Haikou, China
E-mail: zhbahn@vip.qq.com

***Abstract*—Nowadays with the rapid development of network-based services and users of the internet in everyday life, intrusion detection becomes a promising area of research in the domain of security. Intrusion detection system (IDS) can detect the intrusions of someone who is not authorized to the present computer system automatically, so intrusion detection system has emerged as an essential component and an important technique for network security.**

Extreme learning machine (ELM) is an interested area of research for detecting possible intrusions and attacks. In this paper, we propose an improved learning algorithm named self-adaptive differential evolution extreme learning machine (SADE-ELM) for classifying and detecting the intrusions. We compare our methods with commonly used ELM, DE-ELM techniques in classifications. Simulation results show that the proposed SADE-ELM approach achieves higher detection accuracy in classification case.

Keywords-Extreme learning machines; Differential evolution extreme learning machines; Self-adaptive differential evolution extreme learning machines; Intrusion detection; Network security

I. INTRODUCTION

Intrusion into computer networks and systems is a major threat in today’s network centric world. Few most prevalent intrusion attacks include Denial-of-Service (DoS) attacks, Distributed- Denial-of-Service (DDoS) at-tacks, probing based attacks and account takeover attacks. Intrusion detection identifies computer attacks by observing various records processed on the network. Intrusion detection models are classified into two variants, misuse detection and anomaly detection systems. Misuse detection can discover intrusions based on a known pattern also known as signatures [1]. Anomaly detection can identify the malicious activities by observing the deviation from normal network traffic pattern [2]. Hence anomaly detection can identify new anomalies. The difficulty with the current developmental techniques is the high false positive rate and low false negative rate.

Recently, a new fast learning neural algorithm for SLFNs, named extreme learning machine (ELM) [3,4], was devel-oped to improve the efficiency of SLFNs. Different from the conventional learning algorithms for neural networks (such as BP algorithms[5]), which may face difficulties in manually tuning control parameters (learning rate, learn-ing epochs, etc.) and/or local minima, ELM is fully auto-matically implemented without iterative tuning, and in theory, no intervention is required from users. Further-more, It was popular for its fast training speed by means of utilizing random hidden node parameters and calculating the output weights with least square algorithm [6-10]. However, in ELM, the number of hidden nodes is assigned a priori, the hidden node parameters are randomly chosen and they remain unchanged during the training phase. Many non-optimal nodes may exist and contribute less in minimizing the cost function. Moreover, in [11] Huang et al. pointed out that ELM tends to require more hidden nodes than conventional tuning-based algorithms [12,13] in many cases.

Differential evolution (DE) [14] which is a simple but powerful population-based stochastic direct searching technique is a frequently used method for selecting the network parameters [15–17]. In [15], DE is directly adopt-ed as a training algorithm for feed forward networks where all the network parameters are encoded into one population vector and the error function between the network approximate output and the expected output is used as the fitness function to evaluate all the populations. However, Subudhi and Jena [16] have pointed out that using the DE approach alone for the network training may yield a slow convergence speed. Therefore, in [17], a new algorithm named evolutionary extreme learning machine (DE-ELM) based on DE and ELM has been developed for SLFNs. Using the DE method to optimize the network input parameters and the ELM algorithm to calculate the network output weights, DE-ELM has shown several promising features. It not only ensures a more compact network size than ELM, but also has better generalization performance.

However, in the above DE based neural network training algorithms, the trial vector generation strategies and the control parameters in DE have to be manually chosen. For example, the control parameters in DE-ELM are manually selected according to an empirical suggestion and the simple random generation method is adopted to produce the trial vector. As pointed out by many researchers, the performance of the DE algorithm highly depends on the chosen trial vector generation strategy and the control parameters, and inappropriate choices of strategies and control parameters may result in premature convergence or stagnation. Therefore, we propose a novel learning algorithm named self-adaptive evolutionary extreme learning machine (SaDE-ELM) for SLFNs. In SaDE-ELM, the hidden node learning parameters are optimized by the self-adaptive differential evolution algorithm. We verify our approach using the data originated from the 1998 DARPA Intrusion Detection Evaluation Program 1999, which is adopted in the Data Mining and Knowledge Discovery (KDD) competition [18], and considered as a common benchmark for evaluating intrusion detection techniques [19-21]. In this benchmark, there are four types of attacks, Denial of Service (DoS) attack, user to root attack, remote to user attack and probing attack. A denial of service attack is an attempt to make a computer resource unavailable or respond slowly to its legitimate users. User to root attack basically tries to exploit vulnerability to gain root access to the system. Remote to user attack is that attackers remotely exploit vulnerability of a machine to gain local access as a user. Probing are attacks that are trying to access computers, computer systems, networks or applications for weakness. In the following, we first review common methods for intrusion detection and classification.

The rest of the paper is organized as follows. In section II, a brief introduction to ELM and SaDE are given. In Section III, we introduce model of proposed SaDE-ELM algorithm in detail. In Section IV, we present the dataset we use in our numerical studies and our intrusion detection approach. Experiments for detecting intrusion in network traffic data and performance comparisons between ELM-based techniques and DE-ELM-based techniques are presented in section. In section V, we conclude and summarize our results.

II. BACKGROUND

As a novel training algorithm for SLFNs, ELM is very efficient and effective. In this section, we will give a brief review of ELM. In this section, we briefly review ELM and SaDE-ELM approach for ELM is the foundation of SaDE-ELM.

A. *Extreme learning machine (ELM)*

For *N* arbitrary distinct samples $(\mathrm{x}_j, \mathrm{t}_j)$, where $\mathrm{x}_j = [x_{j1}, x_{j2}, \cdots, x_{jn}]^T \in \mathbb{R}^n$, $\mathrm{t}_j = [t_{j1}, t_{j2}, \cdots, t_{jm}]^T \in \mathbb{R}^m$,

SLFNs with *L* hidden nodes and activation function *g*(*x*) are

$$\sum_{i=1}^{L} \beta_i \, \mathrm{g}_i(\mathrm{x}_j) = \sum_{i=1}^{L} \beta_i \, \mathrm{g}_i(\mathrm{w}_i \cdot \mathrm{x}_j + b_j) = \mathrm{o}_j \qquad (j = 1, 2, ..., N) \tag{1}$$

where $\mathbf{w}_i$= [$w_{\mathrm{i}1}$, $w_{\mathrm{i}2}$, ..., w_{in}]*T* is the weight vector connecting the *i*th hidden node and the input nodes, β_i = [$\beta_{\mathrm{i}1}$, $\beta_{\mathrm{i}2}$, ..., β_{im}]^T is the weight vector connecting the *i*th hidden node and the output nodes, b_i is the threshold of the *i*th hidden node, $\mathrm{w}_i \cdot \mathrm{x}_j$ denotes the inner product of w$_i$ and x$_j$, *g*(*x*) is activation function and Sigmoid, Sine, Hardlim and other functions are commonly used. The output nodes are chosen linear in this paper, and o$_j$ = [$o_{\mathrm{j}1}$, $o_{\mathrm{j}2}$, ..., o_{jm}]^T is the *j*th output vector of the SLFNs [22].

The SLFNs with *L* hidden nodes and activation function g(x) can approximate these N samples with zero error. It means $\sum_{j=1}^{L} \left\| \mathrm{o}_j - \mathrm{t}_j \right\| = 0$ and there exist β_i, $\mathbf{w}_i$ and b_i such that

$$\sum_{i=1}^{L} \beta_i \, \mathrm{g}_i(\mathrm{x}_j) = \sum_{i=1}^{L} \beta_i \, \mathrm{g}_i(\mathrm{w}_i \cdot \mathrm{x}_j + b_j) = \mathrm{t}_j \qquad (j = 1, 2, ..., N) \tag{2}$$

The equation above can be expressed compactly as follows:

$$\mathbf{H}\circledR = \mathbf{T} \tag{3}$$

Where

$$\mathrm{H}(\mathrm{w}_1, \mathrm{w}_2, \cdots, \mathrm{w}_L, b_1, b_2, \cdots, b_L, \mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_L)$$

$$=[h_{ij}]= \begin{bmatrix} \mathrm{g}(\mathrm{w}_1 \cdot \mathrm{x}_1 + b_1) & \mathrm{g}(\mathrm{w}_1 \cdot \mathrm{x}_1 + b_2) & \cdots & \mathrm{g}(\mathrm{w}_1 \cdot \mathrm{x}_1 + b_L) \\ \mathrm{g}(\mathrm{w}_1 \cdot \mathrm{x}_2 + b_1) & \mathrm{g}(\mathrm{w}_2 \cdot \mathrm{x}_2 + b_2) & \cdots & \mathrm{g}(\mathrm{w}_L \cdot \mathrm{x}_2 + b_L) \\ \vdots & \vdots & & \vdots \\ \mathrm{g}(\mathrm{w}_1 \cdot \mathrm{x}_N + b_1) & \mathrm{g}(\mathrm{w}_2 \cdot \mathrm{x}_N + b_2) & \cdots & \mathrm{g}(\mathrm{w}_L \cdot \mathrm{x}_N + b_L) \end{bmatrix}_{N \times L}$$

$$\beta = \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{L1} & \beta_{L2} & \cdots & \beta_{Lm} \end{pmatrix} \quad \text{and} \quad \mathrm{T} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ t_{N1} & t_{N2} & \cdots & t_{Nm} \end{pmatrix}$$

The matric H is called the hidden layer output matrix of the neural network and the ith column of H is the ith hidden node output with respect to inputs x1 , x2, ..., xN .

By simply randomly choosing hidden nodes and then adjusting the output weights, single hidden layer feedforward networks (SLFNs) work as universal approximators with any bounded non-linear piecewise continuous functions for additive nodes [23]. ELM algorithm claims that the hidden node parameters can be randomly assigned [3,4], then the system equation becomes a linear model and the network output weights can be analytically determined by finding a least-square solution of this linear system as follow

$$\widehat{\beta} = \mathbf{H}^{\dagger}\mathbf{T} \tag{4}$$

Where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse of matrix **H**. Then the output function of ELM can be modeled as follows.

$$\phi(\boldsymbol{\xi}) = \boldsymbol{\gamma}(\boldsymbol{\xi})® = \boldsymbol{\gamma}(\boldsymbol{\xi})\mathbf{H}□\mathbf{T} \tag{5}$$

Moreover, it should be noted that many nonlinear activation and kernel functions can be used in ELM.

B. Self-adaptive differential evolution

Differential evolution (DE), proposed by Storn and Price in 1995, is a simple yet powerful evolutionary algorithm (EA) [24]. There are three parameters in DE algorithm, which are the population size *NP*, mutation scaling factor *F* and crossover rate *CR*. *NP* is a problem-dependent parameter, while *F* and *CR* are very sensitive to the performance at different stages of evolution. To overcome the limitations of choosing the parameters in DE, Brest et al. [25] proposed a parameter adaptation technique to choose the mutation scaling factor *F* and crossover rate *CR* namely SADE-ELM algorithm which performs better than the basic DE algorithm. In general, SADE algorithm is composed of three main steps: mutation, crossover, and selection [26].

We consider the following optimization problem:

Minimize $f(x_i)$, $x_i \in R_D$

where $x_i = [x_{i1}, x_{i2}, \cdots, x_{iD}]^T, i = 1, 2, \cdots, NP$ is a target vector of D decision variables. During the mutation operation, mutant vector v_i is generated by mutation strategy in the current population:

$$v_i = x_{r1} + F \cdot (x_{r2} - x_{r3}) \tag{6}$$

where $r1, r2, r3$ are mutually exclusive integers randomly chosen in the range [1,*NP*], and $r1 \neq r2 \neq r3 \neq i$.

Following mutation, trial vector u_i is generated between x_i and v_i during crossover operation where the most widely used operator is the binomial crossover performed as follows:

$$u_{ij} = \begin{cases} v_{ij}, & if\,(\text{rndreal}(0,1) < \text{CR} \;\text{ or }\; j = j_{\text{rand}}), \\ x_{ij}, & otherwise \end{cases} \tag{7}$$

Where j_{rand} is a integer randomly chosen in the range [1, *D*], and rndreal(0, 1) is a real number randomly generated in (0, 1). Finally, to keep the population size constant during the evolution, the selection operation is used to determine whether the trial or the target vector survives to the next generation according to one-to-one selection:

$$x_i = \begin{cases} u_{i,} & if\,(\,f(u_i)\,f(x_i)) \\ x_i, & otherwise \end{cases} \tag{8}$$

Where $f(x)$ is the optimized objective function. During the evolution, *F* and *CR* are adaptively tuned to improve the performance of DE for each individual

$$F_{i,G+1} = \begin{cases} F_l + rand_1 \cdot F_u & if\,(rand_2 < \tau_1) \\ F_{i,G} & \text{otherwise} \end{cases} \tag{9}$$

$$CR_{i,G+1} = \begin{cases} rand_3 & if\,(rand_4 < \tau_2) \\ CR_{i,G} & \text{otherwise} \end{cases} \tag{10}$$

Where $F_{i;G+1}$ and $CR_{i;G+1}$ are the mutation scaling factor and crossover rate for *i* individual in *G* generation respectively, *randj*=1*;*2*;*3*;*4 are randomly chosen from (0, 1), τ_1 and τ_2 both valued 0.1 which is used to control the generation of *F* and *CR*, *Fl* valued 0.1 and *Fu* is valued 0.9. In the first generation, *F* and *CR* are initialized to 0.5.

III. Model of Proposed SADE-ELM Algorithm

Since the ELM generates the input weights and hidden biases arbitrarily which are the basic of calculating the output weights, it may not reach the optimal result in classification or regression. Thus, a hybrid approach integrated self-adaptive differential evolution algorithm and extreme learning machine namely SADE-ELM algorithm to optimize the input weights and hidden biases is able to obtain better generalization performance than ELM algorithm [17].

In SaDE-ELM, we proposed SaDE-ELM for SLFNs by incorporating the self-adaptive differential evolution algorithm [25] to optimize the network input weights and hidden node biases and the extreme learning machine to derive the network output weights.

Given a set of training data and L hidden nodes with an activation function g(·), we summarize the SaDE-ELM algorithm in the following steps.

Step 1. Initialization

A set of *NP* vectors where each one includes all the network hidden node parameters are initialized as the populations of the first generation

$$\theta_{k,G} = \left[\mathrm{w}_{1,k,G}^T, \cdots, \mathrm{w}_{L,k,G}^T, b_{1,k,G}^T, \cdots, b_{L,k,G}^T\right] \tag{11}$$

where w_j and b_j (j = 1, … , *L*) are randomly generated, *G* represents the generation and *k* = 1, 2,…, *NP*.

Step 2. Calculations of output weights and RMSE

Calculate the network output weight matrix and root mean square error (RMSE) with respect to each population vector with the following equations, respectively.

$$\beta_{k,G} = \mathrm{H}^{\dagger}_{k,G}\mathrm{T} \tag{13}$$

$$\mathrm{RMSE}_{k,G} = \sqrt{\frac{\sum_{i=1}^{N}\left\|\sum_{j=1}^{L}\beta_j g(\mathrm{w}_{j,k,G}, b_{j,k,G}, x_i) - t_i\right\|}{m \times N}} \tag{14}$$

Then use the value of RMSE to calculate the new best population vector $\theta_{k,G+1}$ with the following equation.

$$\theta_{k,G+1} = \begin{cases} u_{k,G+1} & if\ (\mathrm{RMSE}_{\theta_{k,G}} \text{-} \mathrm{RMSE}_{u_{k,G+1}}) > \varepsilon \cdot \mathrm{RMSE}_{\theta_{k,G}} \\ u_{k,G+1} & if\ \left|\mathrm{RMSE}_{\theta_{k,G}} \text{-} \mathrm{RMSE}_{u_{k,G+1}}\right| < \varepsilon \cdot \mathrm{RMSE}_{\theta_{k,G}} \\ & \quad and \left\|\beta_{u_{k,G+1}}\right\| < \left\|\beta_{\theta,G}\right\| \\ \theta_{i,G} & \text{otherwise} \end{cases} \tag{15}$$

where ε is the preset small positive tolerance rate. In the first generation, the population vector with the best RMSE is stored as θbest,1 and RMSEθbest,1 .

All the trial vectors uk,G+1 generated at the (G+1)th generation are evaluated using equation(11) .The norm of the output weight $\|\beta\|$ is added as one more criteria for the trial vector selection as pointed out in [23] that the neural networks tend to have better generalization performance with smaller weights.

The three operations mutation, crossover and selection are repeated until the preset goal is met or the maximum learning iterations are completed. At last we calculate the output weigh $\beta = [\beta_{i1}\ \beta_{i2}\ \cdots\ \beta_{iL}]^T$ with equation $\beta = \mathbf{H}^{\dagger}\mathbf{T}$.

IV. INTRUSION DETECTION USING SADE-ELM

In this section, we describe the dataset that we use for our numerical studies, and our SaDE-ELM approach to classification of intrusions in the data.

A. Dataset Description

The dataset we use is from the 1998 DAPRA intrusion detection program. During the evaluation program, an environment was set up in Lincoln Labs to record 9 weeks of raw TCP/IP dump data for a network simulating a typical U.S. air force LAN. Then the LAN was operated under a real environment and blasted with multiple attacks. After that, 7 weeks of raw tcpdump data was processed into millions of connection records. Finally, 41 quantitative and qualitative features were extracted using data mining techniques. The detail of the feature extraction can be found in [27].

Four main categories of attacks were simulated:

1) DoS: denial of service attack

2) R2L: unauthorized access from a remote machine

3) U2R: unauthorized access to local root previledges

4) Probing: surveillance and other probing

In the intrusion detection simulation, the dataset was labeled with 22 attack types falling into the four categories shown in Table I. The feature list and its descriptions are in Tables II, III and IV.

TABLE I. TABLE I ATTACK TYPE

Denial of Service	**User to Root**	**Remote to User**	**Probing**
Back Neptune Land Teardrop Ping of Death Smurf	Perl Buffer Overflow Load Module Rootkit	FTP Write Guess Password Imap Multihop Phf Spy Warezclient Warezmaster	IP Sweep Nmap Port Sweep Satan

B. Intrusion Detection System using SaDE-ELM

Our SaDE-ELM intrusion detection method has the following steps. We also use a ELM method to classify the data to provide a comparison benchmark.

Step 1. Data pre-processing: a data processing script is used to convert the raw TCP/IP dump data into machine readable form.

Step 2. Training phase: SaDE-ELM and ELM are trained on normal data and different types of attacks. For the binary classification case, the data has 41 features and falls into 2 classes: normal and attack; for the multi-class classification case, the data has 41 features and falls into 23 classes: normal and 22 types of attack. The model is trained in a large program which can test immediately after the training completed. According to SaDE-ELM theory that has been introduced above, we can summarize the following steps.

For N arbitrary distinct samples (xi,ti), i = 1,... N,and hidden nodes and activation function g(x):

2.1) A set of NP individual parameter vectors $\theta_{k,G}$ (k = 1, 2 . . . NP), where each one includes all the network hidden node parameters are initialized as the populations of the first generation;

2.2) In the case of g(x) and L are invariable run the three operations including mutation, crossover and selection to produce the new population, and the process is repeated until the stop condition is completed.

2.3) Changing the type of g(x) and increase the number of hidden nodes L gradually from one to find the most suitable g(x) and L to construct an optimal forecasting model with the best testing accuracy;

2.4) Calculating the output matrix according to Eq.(4);

2.5) Calculating the output weights $\beta = \mathrm{H}^{\dagger}\mathrm{H}$, where T = [t1, . . . tN] and $\mathbf{H}^{\dagger} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}$.

Step 3. Testing phase: ELM, DE-ELM and SaDE-ELM are used to predict the type of each data point in the testing dataset, and their performances are compared.

Both ELM and SaDE-ELM cannot process symbolic data, so the following method is used to convert symbolic data into continuous data without affecting the performance. As can be seen from the feature description table, there are several symbolic features in the dataset. For features like land, logged in, root shell, is host login and is guest login that take values 0 or 1, so we can handle these features in the same way as continuous features. Other features like protocol, service and flag have more than 2 different values. For example, there are three different values in feature protocol TCP, UDP and ICMP. We represent these three category attributes TCP, UDP, ICMP using (0,0,1), (0,1,0) and (1,0,0). The same method is applied to encode the features service and flag. Experiments have shown that if the number of values in an attribute is not too large, this coding is more stable than using a single number. The simulation of the three algorithms on all datasets are carried out using MATLAB 2013a on a machine with an Intel Core 2 Duo, 2.26GHz CPU and 4GB RAM.

V. SIMULATION RESULTS

The datasets being tested are 2000, 4000, 8000 connection data chosen randomly from the dataset downloaded from the website [18]. We split them equally into training data and testing data. Simulation results including average testing accuracy and corresponding 95% confidence interval are given in Table IV.

In order to test the relationship between SaDE-ELM and the number of hidden layer, according to the different number of hidden layer nodes, we made classification tests using ELM, DE-ELM and SaDE-ELM respectively. Simulation results are given in Table V.

Figure1 and Fig .2 show the time spent by ELM, DE-ELM and SaDE-ELM when training and testing the same size of dataset. It can be seen that the training time and testing time spent by SaDE-ELM increase sharply when the size of data increases. In comparison, ELM and DE-ELM increase slowly when the number of data increases. Eventually, DE-ELM starts consuming more time for both training and testing than ELM.

A clear time consumption comparison can be seen from Fig .1and Fig .2. From the results, we can conclude that ELM performs better than DE-ELM and SaDE-ELM in terms of speed. To increase accuracy, we can implement SaDE-ELM. This shows that our proposed SaDE-ELM methods have better scalability than ELM and DE-ELM when classifying network traffic for intrusion detection.

TABLE II. TABLEII BASIC FEATURES OF INDIVIDUAL TCP CONNECTIONS

feature name	description	type
Duration	length (number of seconds) of the connection	continuous
protocol_type	type of the protocol	discrete discrete
service	network service on the destination	continuous continuous
src_bytes	number of data bytes from source to destination	discrete discrete
dst_bytes	number of data bytes from destination to source	continuous continuous
flag	normal or error status of the connection	
land	1 if connection is from/to the same host/port, 0 otherwise	
wrong_fragment	number of “wrong” fragments	
urgent	number of urgent packets	

TABLE III. TABLEIII CONTENT FEATURES WITHIN A CONNECTION SUGGESTED BY DOMAIN KNOWLEDGE

feature name	description	type
hot	number of “hot” indicators	continuous
num failed logins	number of failed login attempts	continuous
logged in	1 if successfully logged in, 0 otherwise	discrete
num compromised	number of “compromised” conditions	continuous
root shell	1 if root shell is obtained, 0 otherwise	discrete
su attempted	1 if “su root” command attempted, 0 otherwise	discrete
num root	number of “root” accesses	continuous
num file creations	number of file creation operations	continuous
num shells	number of shell prompts	continuous
num access files	number of operations on access control files	continuous
num outbound cmds	number of outbound commands in an ftp session	continuous
is hot login	1 if the login belongs to the “hot” list, 0 otherwise	discrete
is guest login	1 if the login is a “guest”login, 0 otherwise	discrete

TABLE IV. TABLE IV TRAFFIC FEATURES COMPUTED USING A TWO-SECOND TIME WINDOW

feature name	**description**	**type**
count	number of connections to the same host as the current connection in the past two seconds	continuous
serror rate	% of connections that have "SYN" errors	continuous
rerror rate	% of connections that have "REJ" errors	continuous
same srv rate	% of connections to the same service	continuous
diff srv rate	% of connections to different services	continuous
srv count	number of connections to the same service as the current connection in the past two seconds	continuous
srv serror rate	% of connections that have "SYN" errors	continuous
srv rerror rate	% of connections that have "REJ" errors	continuous
srv diff host rate	% of connections to different hosts	continuous

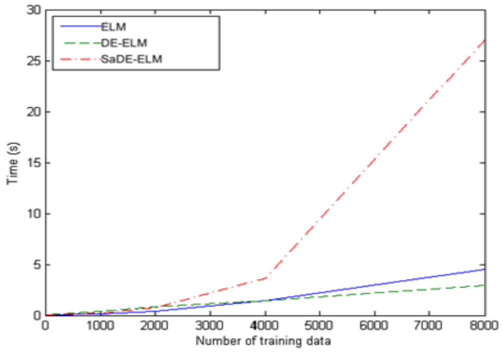


Figure 1. Training time comparison

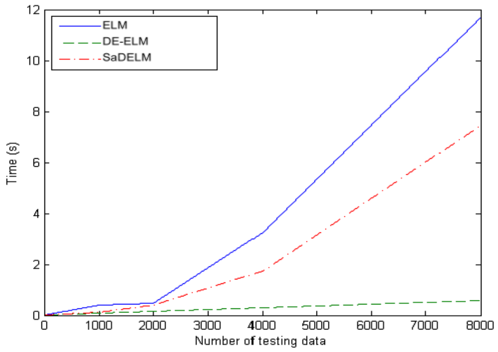


Figure 2. Testing time comparison

TABLE V. TABLE V. PERFORMANCE COMPARISON RESULTS

Dataset Size	**ELM**		**DE-ELM**		**SaDE-ELM**	
Training/Testing	Accuracy (%)	95% Confidence Interval (%)	Accuracy (%)	95% Confidence Interval (%)	Accuracy (%)	95% Confidence Interval
1000/1000	99.32	99.08 - 99.47	99.33	99.15 - 99.51	99.55	99.05 - 99.65
2000/2000	99.10	98.82 - 99.23	99.24	98.90 - 99.44	99.47	99.25 - 98.58
4000/4000	99.07	98.79 - 9.28	99.18	99.01 - 99.26	99.35	99.11 - 99.65

VI. CONCLUSION

In this paper, we have made a comparison by the use of ELM, DE-ELM and SaDE-ELM for intrusion detection in a computer network. For the SaDE-ELM, By incorporating the self-adaptive differential evolution algorithm to optimize the network hidden node parameters and employing the extreme learning machine to derived the network output weights. Obviously, the proposed SaDE-ELM can obtain higher accuracy.

Whether to use ELM, DE-ELM or SaDE-ELM in implementing an intrusion detection system depends on the type of intrusion likely to occur. For example in a DDoS attack, the attacker usually controls thousands of agents to send a large number of TCP SYN packets to a victim's server port. When the port is actively listening for connection requests, the victim would respond by sending back ACK packets. However, the victim will not get further responses and keep the connections half-open, which would eventually quickly consume all the memory allocated for pending connections. The victim's server would then no longer be able to process new requests from normal clients. If we can correctly detect more than 90% of the attack connections and drop these, we can effectively prevent the DDoS attacker from overwhelming the server. For DDoS attack detection, basic ELM with sigmoid additive neurons would be a good choice since it has significantly shorter training times compared to other techniques. On the other hand, attacks like user to root attack exploit the victim's vulnerability to gain root access and may not create as many connections as DDoS attack. Each connection by a successful attack however provides root access to the system. Therefore, in this case, detection accuracy matters more than speed. To detect this kind of attack, SaDE-ELM would be preferred.

ACKNOWLEDGMENT

This research was supported by science research project of Education Department of Hainan province (Hnky2017ZD-20).

REFERENCES

[1] Ilgun, K., Kemmerer, R.A., Porras, P.A., 1995. State transition analysis: a rule-based intrusion detection approach. IEEE Trans. Software Eng. 21 (3), 181–199.

[2] Ikram S T, Cherukuri A K. Improving Accuracy of Intrusion Detection Model Using PCA and optimized SVM[J]. CIT. Journal of Computing and Information Technology, 2016, 24(2): 133-148.

[3] Huang GB, Zhu QY, Siew CK. Extreme learning machine: a new learning scheme of feedforward neural networks. In: Proceedings of international joint conference on neural networks (IJCNN2004), vol 2, no 25–29, pp 985–990.

[4] Huang GB, Zhu QY, Siew CK. Extreme learning machine: theory and applications. Neurocomputing 70(1–3):489–501.

[5] Espana-Boquera S, Zamora-Mart ńez F, Castro-Bleda M J, et al. Efficient BP algorithms for general feedforward neural networks[C]//International Work-Conference on the Interplay Between Natural and Artificial Computation. Springer Berlin Heidelberg, 2007: 327-336.

[6] G. Thatte, U. Mitra, and J. Heidemann, "Parametric methods for anomaly detection in aggregate traffic," IEEE/ACM Transactions on Networking, vol. 19, no. 2, pp. 512–525, April 2011.

[7] M. Qin and K. Hwang, "Frequent episode rules for internet anomaly detection," in Proceedings of the Network Computing and Applications, Third IEEE International Symposium. Washington, DC, USA: IEEE Computer Society, 2004, pp. 161–168.

[8] X. He, C. Papadopoulos, J. Heidemann, U. Mitra, and U. Riaz, "Remote detection of bottleneck links using spectral and statistical methods," Computer Networks, vol. 53, pp. 279–298, February 2009.

[9] W. W. Streilein, R. K. Cunningham, and S. E. Webster, "Improved detec- tion of low-profile probe and denial-of-service attacks," in Proceedings of the 2001 Workshop on Statistical and Machine Learning Techniques in Computer Intrusion Detection, June 2001.

[10] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273–297, 1995.

[11] G.-B. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: a survey," International Journal of Machine Leaning and Cybernetics, vol. 2, no. 2, pp. 107–122, 2011.

[12] G. Tandon, "Weighting versus pruning in rule validation for detecting network and host anomalies," in In Proceedings of the 13th ACM SIGKDD international. ACM Press, 2007.

[13] Y. Liao and V. R. Vemuri, "Use of k-nearest neighbor classifier for intrusion detection," Computers & Security, vol. 25, pp. 439–448, 2002.

[14] Storn R, Price K (2004) Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J Glob Optim 11(4):341–359

[15] Ilonen J, Kamarainen JI, Lampinen J (2003) Differential evolution training algorithm for feedforward neural networks. Neural Process Lett 17:93–105

[16] Subudhi B, Jena D (2008) Differential evolution and levenberg marquardt trained neural network scheme for nonlinear system identification. Neural Process Lett 27:285–296.

[17] Zhu Q-Y, Qin A-K, Suganthan P-N, Huang G-B (2005) Evolutionary extreme learning machine. Pattern Recog 38(10):1759–1763

[18] (1999)KDDCUPdataset. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[19] S. Mukkamala and A. Sung, "Detecting denial of service attacks using support vector machines," in Proceedings of the 12th IEEE International Conference on Fuzzy Systems, 2003.

[20] M. Luo, L. Wang, H. Zhang, and J. Chen, "A research on intrusion detection based on unsupervised clustering and support vector machine," in Information and Communications Security, ser. Lecture Notes in Computer Science, S. Qing, D. Gollmann, and J. Zhou, Eds. Springer Berlin / Heidelberg, 2003, vol. 2836, pp. 325–336.

[21] D. Kim and J. Park, "Network-based intrusion detection with support vector machines," in Information Networking, ser. Lecture Notes in Computer Science, H.-K. Kahng, Ed. Springer Berlin / Heidelberg, 2003, vol. 2662, pp. 747–756.

[22] Lin, Y., Lv, F., Zhu S., Yang, M., Cour, T., Yu, K., Cao, L., Huang, T.S.: Large-scale image classification: fast feature extraction and SVM training. In: Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1689-1696 (2011).

[23] Huang G-B, Chen L, Siew CK (2006) Universal approximation using incremental constructive feedforward networks with random hidden nodes. IEEE Trans Neural Netw17(4):879-892

[24] R. Storn, K. Price, Differential evolution-A simple and efficient heuristic for global optimization over continuous spaces, Journal of Global Optimization, 11 (1997) 341-359.

[25] J. Brest, S. Greiner, B. Boˇskovi´c, M. Mernik, V. ˇZumer, Self-adapting control parameters in differential evolution: A comprehensive study on numerical benchmark problems. IEEE Transactions on Evolutionary Computation, 10 (2006) 646-657.

[26] J. Wu, Z. H. Cai, Attribute Weighting via Differential Evolution Algorithm for Attribute Weighted Naive Bayes (WNB), Journal of Computational Information Systems, 7 (2011) 1672-1679.

[27] S. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. K. Chan, "Costbased modeling for fraud and intrusion detection: results from the JAM project," in Proceedings of DARPA Information Survivability Conference and Exposition, vol. 2, January 2002, pp. 130‐144..

Job to Major (J2M): an Open Source Based Application

Haitao Liu
Department of Information Management
Dalian Neusoft University of Information
Dalian, China
E-mail: liuhaitao@neusoft.edu.cn

Jiacong Zhao*
Department of Information Management
Dalian Neusoft University of Information
Dalian, China
E-mail: zhaojiacong@neusoft.edu.cn
*The corresponding author

Dongzhao Zhou
Department of Information Management
Dalian Neusoft University of Information
Dalian, China
E-mail: zhoudongzhao@neusoft.edu.cn

Shuang Gao
Department of Information Management
Dalian Neusoft University of Information
Dalian, China
E-mail: gaoshuang@neusoft.edu.cn

***Abstract*—This paper presents Job to Major (J2M), an open-source tool design to link job requirements with University of Southampton's major information. J2M currently provides two functions: suggesting the most suitable majors to potential students of the university based on the job requirements, find jobs that are closely link to what students acquire from their major. The development of J2M is according to Garrett's model which divides the application development mainly into three stages business goals analysis, system design and implementation. J2M build on existing open standards and supported by open-source of University of Southampton, Universal JobMatch provided by gov.uk and additional tools. The evaluation of J2M tries to show the functions achievement of J2M.**

Keywords- Job to Major (J2M); Open innovation; Business; Design; Implementation

I. INTRODUCTION

Currently, both students and universities are placed in dilemma in terms of the relatedness of college major and job market. Firstly, students attend college and select degree fields in the hope of succeeding in the labor market [1]. Furthermore, the ability to utilize the investment in schooling in future employment [1,2,3] is one aspect of labor market success. Hence, most potential students chose their majors based on requirements of the job. However, it is difficult for them to compare which majors of the university contribute most to their future career through general information searching. In addition, occupation specific skills related to the current occupation increase wages [2]. In contrast, wages are lower for mismatched workers who are working in a job that uses fewer of the occupation specific skills learned by graduates in the major. This dilemma is described as mismatch between education and job market [1]. Thus, in order to make what they have learnt valuable and get relative higher salary, graduates try to find a job that highly corresponds to what they have learned. Secondly, it is unreality to require universities to keep their majors' content up to date with the rapidly changing job market requirement based on traditional access to job information. Current solutions to students' career development can mainly be divided into three types. One is web applications like Total-Jobs and Target-Jobs are only focus on job finding. One is application like C4S, it only pays attention to major finding and the application is not user-friendly. The third way is the employment center of university which is both time and cost wasting. Thus, in order to address these problems, a new application should be proposed to link job requirements with students major information.

Open innovation allows organizations to look beyond their internal resources to develop new products, services, and sources of revenue, so as to solve real world problems [3]. Additionally, open data can be accessed and used freely by anyone and open data innovations usually contribute more opportunities to digitize, automate and optimize many kinds of current services[4]. In order to benefit to students, universities, and even parents and employers, the proposed application will based on the open source of job requirements and college majors' information. Therefore, we propose a open-source web application Job to Major (J2M) to solve the mentioned dilemma for students of University of Southampton. Open data innovation project delivered a rich and varied program of activities, like business, technology and even ethics. Thus, the development of J2M was organized into five sections. The following sections firstly demonstrate the theoretical background of open data innovation. Secondly, specifying the design. Then, turning to the implementation of J2M and following the texting of the whole system. Finally, evaluating the application through business feasibility, social contribution and techniques aspects.

II. THEORETICAL BACKGROUND

As the development of information technology, every company is exposed to a widely distributed knowledge world. Companies cannot afford to rely entirely on their own research, in this situation, many innovative firms have shifted to an ‘open innovation’ model, using a wide range of external actors and sources to help them achieve and sustain innovation [4]. In this way, companies can reduce cost for conducting research and development, improve development productivity, increase the accuracy for customer targeting [5]. Additionally, the external resources generally gain by collaborating with local governments, universities, business support services, and other public bodies [6]. It indicates that open innovation businesses often shoulder social responsibilities. However, there are still existing kinds of problems in this process. For example, revealing information not intended for sharing and revealing intellectual property of hosting organization. In order to reduce these risks, how to exploit openness for firms' benefits has been the heart of recent research on innovation [4, 7, 8]. In this situation, open innovation models are proposed to formal the whole development process which includes idea generation, business analysis, application design, application implementation, testing and evaluation. For example, Garrett's model [9] focuses on user experience achievement. It divides the innovation process into five stages: surface, skeleton, structure, scope and strategy. These stages clearly define what goals the innovation try to gain, what tasks should do to support goals and what tools should be used to put tasks into practice. Fugle model [5] divides this process into seven stages, compared with Garrett's model, Fugle focuses more on concept establish and specifics platforming criteria for technical level [10]. It is therefore can be seen that current open innovation model can support the open-based web application development. J2M is developed based on Garrett's model and also considers Fugle model's platforming criteria.

The development of Web innovation applications should also be supported by robust web techniques. Majority of people spend a lot of time on Web. This motivates companies take diverse number of actions for open innovation [11]. For example, Procter & Gamble developed its “Connect + Develop website” to get in touch with external innovators, so that they could contribute through propositions of solutions to P&G problems. OpenCalais is also a kind of open Web service to support text analysis for J2M. With the development of web tools, Web has become a platform for collaboration which triggered the growth of open innovation platforms. These platforms try to leverage the Web technology and most notably its social aspects to help web innovation [11], like Node.JS and Express.JS. Current research also addresses paradigms in open innovation processes, on which we will base our further analysis (Table.1). These paradigms are applied to different research situations to deal with unexplored aspect and consequences in the innovation process. Furthermore, there are many identified Web technologies that are likely to be useful in problem solving processes on the open innovation platforms, like expert finding [12], semantic keyword matching [13, 14] and social propagation [15]. J2M mainly relies on the semantic keyword matching to broad the space of matching possibilities between major information and job requirements. Finally, tools like Sublime Text, SQLite Manager and CSV to JSON Converter are mature enough to deal with objectives like code edit, data storage and data format transformation to support the goals achievement. Previous theoretical background analysis shows that current open innovation model and related web techniques guarantee the development of J2M is practical rather ambitious.

III. DESIGN OF J2M

In order to guarantee the productive development process of a project, it is better to make a plan. Gantt Chart Project Plan will be given firstly. In order to gain high user experience, Garrett's model is introduced for designing J2M, and J2M will be developed through the following three aspects: business strategy, system design and implementation.

TABLE I. TABLE I.OPEN INNOVATION PARADIGMS

Social Behavior on the Web-“Weak Ties” [16]	Explaining the social behavior of people on the Web and to the Open Innovation process.
Cross-Sectorial Problem Solving [17]	Creating solutions to the problem that were at the border or outside companies' area of expertise.
Broadcasting [17]	Representing the distribution of different content related to problems to a dispersed audience

A. Business Analysis

The business goals of J2M is firstly to attract potential students of University of Southampton to use this application to find the suitable major based on the requirements of jobs that they are interested in. Then, attracting current Southampton students to use this application to find a job that highly consistent to what they have learned. In order to guarantee that J2M is customer focused and can survive in the market. Two business models VRINE (Value, Rare, Inimitable& Non-substitutable, Exploitable)[18] and SWOT (Strength, Weakness, Opportunities, Threats) [19] are introduced to analysis the resources and capabilities of J2M. The following five components (Fig .1) are identified important to the sustainable development of this application.

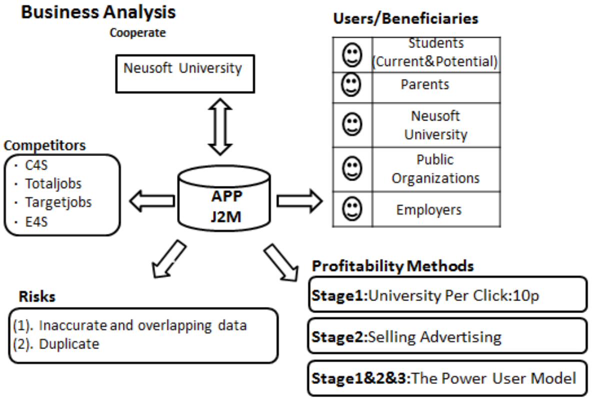


Figure 1. Business Analysis of J2M

1) Target Organization

A starting point for the idea of openness is that a single organization cannot innovate in isolation. It has to engage with different types of partners to acquire ideas and resources from the external environment to stay abreast of competition (Fig .2 [20]) [4, 8]. So we cooperate with Southampton University and develop this application to assist the university in terms of students' career development.

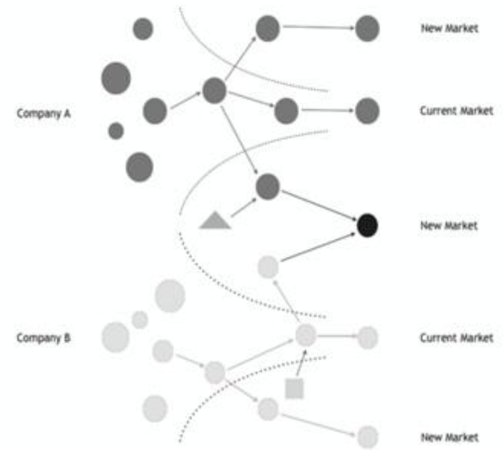


Figure 2. The Open Innovation Process [20]

2) *Customers*

Customer refers to the benefiter of our application. Potential students, based on their dream jobs' requirements, we suggest suitable majors of Southampton University to them. Current students can find jobs that highly match to the acquirements of their majors. Education industry has developed more similar to consumer goods market [21]. University provides its service and reputation to students. This market is also seriously competitive. In order to achievement sustainable development in this market, the university should sensitive to job market. J2M enable the university gain this advantages. Public organizations, like the education department with our data can track the current situation between education and job market. This helps them plan the future education project more practically. Employers usually spend a great deal of time and money to hunt for desired employee. J2M narrows the hunting scope for them, and they just need to pay more attention to specific major students. This is both money and time saving.

3) *Competitors*

Recently, four competitors are identified. Firstly, C4S is limited in major searching and has no fitting function. It is not user-friendly. The Total-Jobs, Target-Jobs and E4S are only for job searching. It can be seen that they all focus on major or job searching but not link them together. Our application fills this gap, which distinguishes us from these competitors.

4) *Risk*

The risk of J2M mainly consists of two parts. One is the dataset itself. The open dataset contains inaccurate and overlapping data, which will easily lead to inaccurate and even wrong output. This risk can be avoided by data refine tools, such as Google Refine. The other kind of risk is mainly from the duplication. Because of the easily access to open source, every can duplicate the proposed idea.

5) *Profitability method*

We design our profitability strategies into three stages:

- **Stage1:** Charge Southampton University by each click.
- **Stage 2:** When we gain relative more users then we can sell some advertising space to the website like C4S and total-jobs.
- **Stage 3:** We firstly provide good but limited information to the users in the previous two stages. Then charge users with the premium service.

B. *J2M Systme Design*

J2M owns two functions: returning a list of possible majors when a job title is input (Fig .3) and returning a list of job titles with description when a major title is selected (Fig .4). To keep implementation works in an appropriate scale, the study program should be limited to Electricity and Computer Science courses in the University of Southampton only. In the same reason, the kinds of job are also filtered to computer and IT related jobs. The detailed structure is as follows.

In Fig .3, when a job title is input to the interface, then it connects to open data Universal JobMatch provided by gov.uk. to filter related jobs, with short job descriptions. Then those job data will be connected to OpenCalaris, the open keyword finder to get keywords for each job. There will be another database in the application, the list of the keywords for each module in the University of Southampton. Those will be already listed up. Then the application will compare the keywords list of each job to the keywords of each module. Finally the list of possible study programs suitable to the job title will be displayed in the interface, with number of hit of keywords.

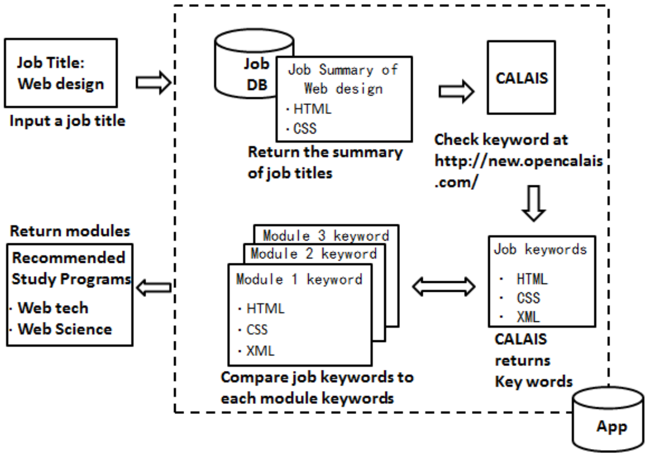
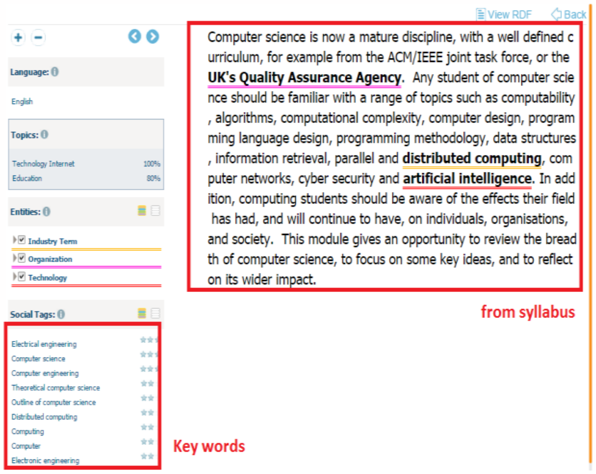


Figure 3. Major Sele cting

In Fig .4, The application works inversely. There will be a dropdown menu of study programs of Electricity and Computer Science in the University of Southampton. When a major is selected, the application will connect to the list of keywords for each module to get temporary keywords. Then it connects to the job database and OpenCalarsi to get keywords match to the temporary keywords, and returns the list of recommended job titles.



Computer science is now a mature discipline, with a well defined curriculum, for example from the ACM/IEEE joint task force, or the **UK's Quality Assurance Agency**. Any student of computer science should be familiar with a range of topics such as computability, algorithms, computational complexity, computer design, programming language design, programming methodology, data structures, information retrieval, parallel and **distributed computing**, computer networks, cyber security and **artificial intelligence**. In addition, computing students should be aware of the effects their field has had, and will continue to have, on individuals, organisations, and society. This module gives an opportunity to review the breadth of computer science, to focus on some key ideas, and to reflect on its wider impact.

Figure 4. Job Selecting

IV. J2M IMPLEMENTATION

The application is modularized. There are modules for connecting to text analysis service, connecting to database to get course module keywords and matching job keywords and course module keywords. Ajax is used for gaining information or refreshing data on the webpage. This can avoid being disturbed by refreshing page and make manipulation more quick and smart. The following part mainly describes the function from job information to module. And the function from module to job is the same.

A. *Frameworks and Development tools*

The frameworks and development tools are the foundation for the application development. To J2M, the frameworks for developing are Node.js and Express.js. Development tools are shown in Table 2.

TABLE II. TABLE II EVELOPMENT TOOLS

Development tools	Function
Sublime text	Code editing
SQLite manager	Data management
CSV to JSON Converter	Convert the open data format into .JSON
Sqlite3	Database for storing keywords of course

B. *Open Data for Job*

Job list is gained from Universal JobMatch website by using API (api.lmiforall.org.uk/api/v1/vacancies/search). Keyword user inputs will be sent through API to search jobs. After gaining job information from API, job titles were picked from the JSON data, and shown in table. Job information of each job is allocated using jQuery click event function. Job description of a specific job can be achieved by clicking job titles.

C. *Open Data for Major*

Firstly visit Website http://data.southampton.ac.uk/dataset/courses.html for major information and download courses.csv file. Then, select module information by filtering out "DEPT_DESC is ECS" and "SUBJ_CODE is COMP" then extract unique PROG_CODE, to get 30 majors related to computer science. For each major, visit university site of module list, for example to get syllabus of all possible modules by http://www.ecs.soton.ac.uk/programmes/msc_computer_science#modules.Fig .5 shows a model example.

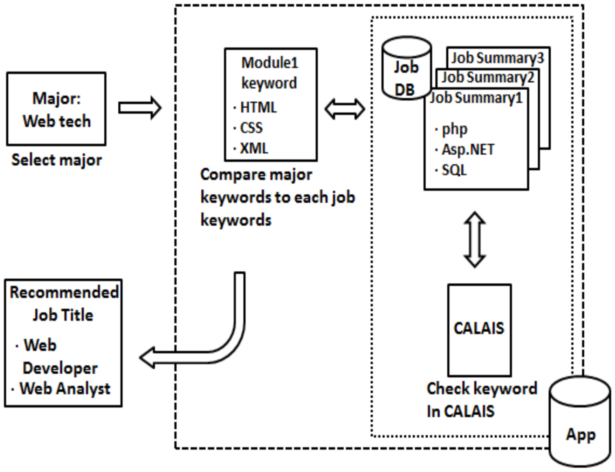
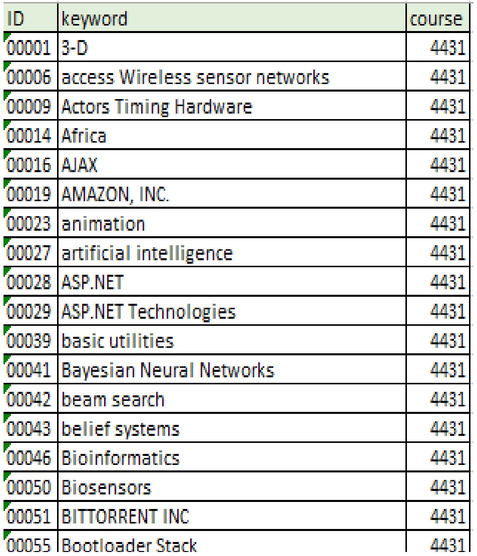


Figure 5. An example of OpenCalais text analysis

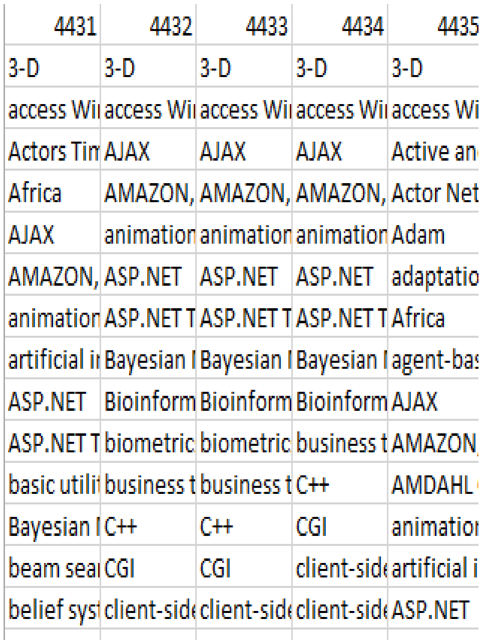
D. *Gaining Keywords*

Job keywords for matching with module keywords are achieved by sending job description to OpenCalais text analysis service. Job description is a part of job information gained from API. OpenCalais API is connected by using Calais node module. Calais node module sends text to analyze and fetches text analysis result. Text is sent by HTTP request using jQuery Ajax. Course module keywords are gained from database. Fig .6 shows an example of OpenCalais text analysis.With OpenCalais the keywords for each model can be achieved, and then list up all keywords for each major (e.g. Fig .7). Then, list up analyzed keywords with major(module) code (e.g. Fig .8).



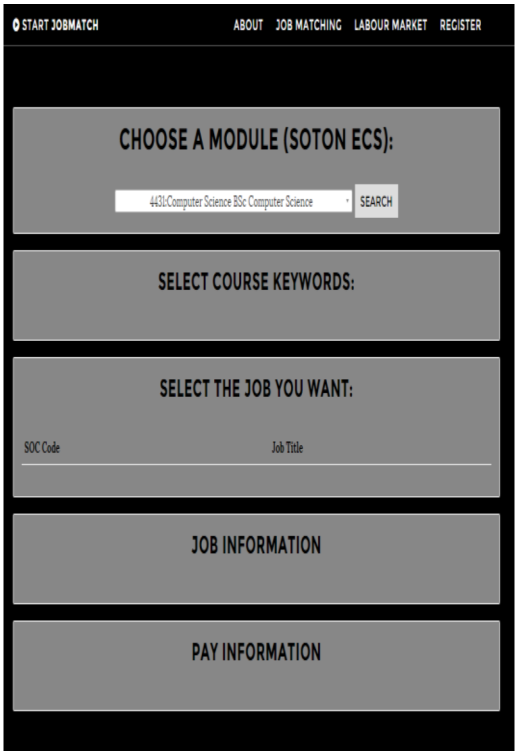
ID	keyword	course
00001	3-D	4431
00006	access Wireless sensor networks	4431
00009	Actors Timing Hardware	4431
00014	Africa	4431
00016	AJAX	4431
00019	AMAZON, INC.	4431
00023	animation	4431
00027	artificial intelligence	4431
00028	ASP.NET	4431
00029	ASP.NET Technologies	4431
00039	basic utilities	4431
00041	Bayesian Neural Networks	4431
00042	beam search	4431
00043	belief systems	4431
00046	Bioinformatics	4431
00050	Biosensors	4431
00051	BITTORRENT INC	4431
00055	Bootloader Stack	4431

Figure 6. An example of module keywords



4431	4432	4433	4434	4435
3-D	3-D	3-D	3-D	3-D
access Wi	access Wi	access Wi	access Wi	access Wi
Actors Tin	AJAX	AJAX	AJAX	Active an
Africa	AMAZON,	AMAZON,	AMAZON,	Actor Net
AJAX	animation	animation	animation	Adam
AMAZON,	ASP.NET	ASP.NET	ASP.NET	adaptatio
animation	ASP.NET T	ASP.NET T	ASP.NET T	Africa
artificial i	Bayesian I	Bayesian I	Bayesian I	agent-bas
ASP.NET	Bioinform	Bioinform	Bioinform	AJAX
ASP.NET T	biometric	biometric	business t	AMAZON,
basic utili	business t	business t	C++	AMDAHL
Bayesian I	C++	C++	CGI	animatio
beam sea	CGI	CGI	client-sid	artificial i
belief sys	client-sid	client-sid	client-sid	ASP.NET

Figure 7. Keywords with major (module) code



4431:Computer Science BSc Computer Science

Figure 8. Figure 8. Job Working

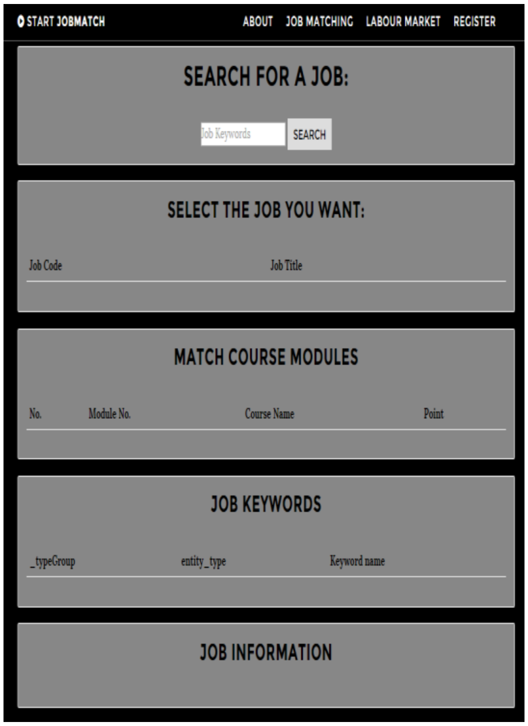


Figure 9. Job Finding

E. Getting Module Recommendation

Module recommendation is made based on the result of keyword matching. Job keywords are sent to matching function module and matched with module keywords. Matching point will be added once keywords match. Matching point calculated in this function will be returned and sorted to recommend the most matched modules.

The testing here is to test whether J2B can achieve functions for job and major selecting.

1) From job requirements to major

When users search job keywords, the “SELECT THE JOB YOU WANT:” part will show you the jobs. After users selecting a job, the rest parts will show the matched courses with points, job keywords and job information.

2) From major to job position

This part is to test whether J2M can find a job position based on the major information (Fig . 9).

The test result shows that J2M can achieve these tow functions well.

V. J2M EVALUATION

The evaluation of J2M will be discussed through two aspects. One is for the evaluation for a web application which should achieve high level of design schema, information content and ease-of-use. The other aspect is for open innovation evaluation. This refers to compared with traditional close data application J2M brings what kinds of discovery and divergence.

A. Web Application Evaluation

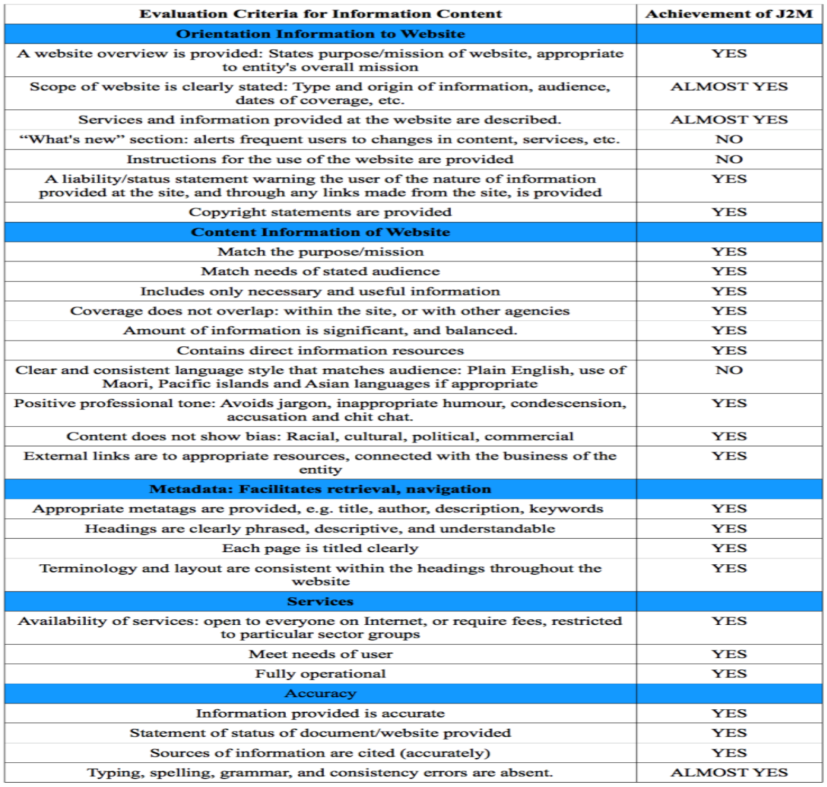
1) Design schema analysis

This part only focuses on static descriptions of the application, which mainly verifies the correctness and consistency of system design specifications [22,23], to enhance the quality of conceptual schemas by looking for design inconstancies and irregularities in the application of design patterns [24]. The patterns addressed by J2M consist of compositions of hypertext elements: pages, units, operations and links. These elements are typically serving application purpose. The testing part shows that J2M can arrange of pages, units, and links for supporting the navigation between job and major information. And in J2M a core object can be accessed via one or more access objects. For example, a job position can be directly accessed via module content key words or from module number which needs more accesses. Hence, J2M achieves acceptable design schema.

2) Information content evaluation

J2M's information will be evaluated by criteria provided by [24]. Five types of criteria are discussed: orientation information to Website, Content Information, Metadata, Services, Accuracy. Table.3 clearly shows that J2M achieves most of the criteria in the information contents.

TABLE III. INFORMATION CONTENT EVALUATION CRITERIA AND J2M ACHIEVEMENT



Evaluation Criteria for Information Content	Achievement of J2M
Orientation Information to Website	
A website overview is provided: States purpose/mission of website, appropriate to entity's overall mission	YES
Scope of website is clearly stated: Type and origin of information, audience, dates of coverage, etc.	ALMOST YES
Services and information provided at the website are described.	ALMOST YES
“What's new” section: alerts frequent users to changes in content, services, etc.	NO
Instructions for the use of the website are provided	NO
A liability/status statement warning the user of the nature of information provided at the site, and through any links made from the site, is provided	YES
Copyright statements are provided	YES
Content Information of Website	
Match the purpose/mission	YES
Match needs of stated audience	YES
Includes only necessary and useful information	YES
Coverage does not overlap: within the site, or with other agencies	YES
Amount of information is significant, and balanced.	YES
Contains direct information resources	YES
Clear and consistent language style that matches audience: Plain English, use of Maori, Pacific islands and Asian languages if appropriate	NO
Positive professional tone: Avoids jargon, inappropriate humour, condescension, accusation and chit chat.	YES
Content does not show bias: Racial, cultural, political, commercial	YES
External links are to appropriate resources, connected with the business of the entity	YES
Metadata: Facilitates retrieval, navigation	
Appropriate metatags are provided, e.g. title, author, description, keywords	YES
Headings are clearly phrased, descriptive, and understandable	YES
Each page is titled clearly	YES
Terminology and layout are consistent within the headings throughout the website	YES
Services	
Availability of services: open to everyone on Internet, or require fees, restricted to particular sector groups	YES
Meet needs of user	YES
Fully operational	YES
Accuracy	
Information provided is accurate	YES
Statement of status of document/website provided	YES
Sources of information are cited (accurately)	YES
Typing, spelling, grammar, and consistency errors are absent.	ALMOST YES

3) *Web usage analysis*

Web usage analysis refers to analyze dynamic data that collected at runtime and produce quality reports on content access and navigation sequences [23]. To J2M usage analysis, links, feedback, accessibility and navigability based on the criteria provided by [22, 23, 24] will be evaluated It can be seen that although J2M can support the job or major searching. But it should be update to a more user-friendly application.

B. *Innovation Evaluation*

The innovation of J2M will be evaluated based on the two fields ground by Garrett's model: business goals and technology. Firstly, traditional close data based business usually benefit to the business conducting company by exploiting customers. Compared with based business, the goals of open business is not only benefit to business conducting companies, but more customer and social responsibility focused. This means that open business tries to establish a win-win commercial environment, where every participant is beneficiary. For example, J2M benefits to a wide group of people as well as itself. Secondly, in terms of technology, one of the significant contributions is concurrency. Unlike closed data application, J2M does not

Need to update its data set frequently. Because, the dataset of J2M is concurrently with job and major information. However, disadvantages exit as well. For example, open innovation should choose the open source seriously, to avoid the inaccurate and fault outcomes caused by the original dataset.

The previous analysis shows that in technology field J2M achieves most of the evaluation criteria. What J2M should improve is to be more user-friendly. As an open data application J2M enjoys both business and technology advantages. However, J2M still needs to pay attention to risks for open data application.

REFERENCES

[1] J. Robst, “Education and job match: The relatedness of college major and work,”Economics of Education Review, vol. 26, 2007, p. 397407.

[2] Bauer, T. K., Educational mismatch and wages: A panel analysis. Economics of Education Review, 21(3), 221–229, 2002

[3] Shaw, Kathryn L. "A formulation of the earnings function using the concept of occupational investment."Journal of Human Resources (1984): 319-340.

[4] Chesbrough, H. W. (2003). Open innovation: The new imperative for creating and profiting from technology. Harvard Business Press.

[5] Du Preez, N; Louw, L. 2008. A Framework for Managing the Innovation Process, Picmet 2008.

[6] Enkel, Ellen, Oliver Gassmann, and Henry Chesbrough. "Open R&D and open innovation: exploring the phenomenon." R&d Management 39.4 (2009): 311-316.

[7] Laursen, K., & Salter, A. (2006). Open for innovation: the role of openness in explaining innovation performance among UK manufacturing firms. Strategic management journal, 27(2), 131-150.

[8] Dahlander, L., & Gann, D. M. (2010). How open is innovation?. Research policy, 39(6), 699-709.

[9] Garrett, Jesse James. "The Elements of User Experience." Jjg. Net (2004).

[10] Marais, S. J., and C. S. L. Schutte. "The development of open innovation models to assist the innovation process." 23rd Annual SAIIE Conference Conference Proceedings. 2009.

[11] Jesic, D., Kovacevic, J., & Stankovic, M. (2011, June). Web technologies for open innovation. In Proceedings of the 3rd International Web Science Conference (p. 20). ACM.

[12] Aleman-Meza, B., Bojars, U., Boley, H., Breslin, J. G., Mochol, M., Polleres, A., et al. 2007. Combining RDF vocabularies for expert finding. Lecture Notes in Computer Science, 4519, 235. Springer. Retrieved from http://www.springerlink.com/index/p6u10781711xp102.pdf.

[13] Ziegler, C.-N., Simon, K., and Lausen, G. 2006. Automatic Computation of Semantic Proximity Using Taxonomic Knowledge Categories and Subject Descriptors. CIKM ’06 Proceedings of the 15th ACM international conference on Information and knowledge management (pp. 465-474). Arlington, Virginia, USA:ACM New York, NY, USA. Retrieved from http://doi.acm.org/10.1145/1183614.1183682.

[14] Cilibrasi, R. L., and Vitanyi, P. M. B. 2007. The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering, 19(3), 370-383. doi: 10.1109/TKDE.2007.48.

[15] Hastings, G. 2007. Social Marketing - Why should the devil have all the best tunes? Elsevier Ltd.

[16] Granovetter, M. 1983. The strength of weak ties: a network theory revisited. Sociological Theory, Volume 1 (1983), 201- 233.

[17] Aleman-Meza, B., Bojars, U., Boley, H., Breslin, J. G., Mochol, M., Polleres, A., et al. 2007. Combining RDF vocabularies for expert finding. Lecture Notes in Computer Science, 4519, 235. Springer. Retrieved from http://www.springerlink.com/index/p6u10781711xp102.pdf.

[18] Carmeli, Abraham. "Assessing core intangible resources." European Management Journal 22.1 (2004): 110-122.

[19] Hill, Terry, and Roy Westbrook. "SWOT analysis: it's time for a product recall."Long range planning 30.1 (1997): 46-52.

[20] Laursen, K., & Salter, A. (2006). Open for innovation: the role of openness in explaining innovation performance among UK manufacturing firms. Strategic management journal, 27(2), 131-150.

[21] Singh, M., “E-Services and Their Role in B2C E-Commerce’, Journal of Managing Service Quality, 12:2, pp.434 – 446, 2002.

[22] S. Comai, M. Matera, and A. Maurino. “A Model and an XSL Framework for Analysing the Quality of WebML Conceptual Schemas”. Proc. of IWCMQ'02 - ER'02 International Workshop on Conceptual Modeling Quality, Tampere, Finland, October 2002.

[23] P. Fraternali, M. Matera, and A. Maurino. “WQA: an XSL Framework for Analyzing the Quality of Web Applications”. Proc. of IWWOST'02 – ECOOP’02 International Workshop on Web Oriented Software Technologies, Malaga, Spain, June 2002.

[24] Fraternali, Piero, Maristella Matera, and Andrea Maurino. "Conceptual-level log analysis for the evaluation of web application quality." Web Congress, 2003. Proceedings. First Latin American. IEEE, 2

The Mining Algorithm of Frequent Itemsets based on Mapreduce and FP-tree

Bo He
School of Computer Science and Engineering
ChongQing University of Technology
400054 ChongQing China
E-mail: hebo@cqut.edu.cn

Hongyuan Zhang
School of Computer Science and Engineering
ChongQing University of Technology
400054 ChongQing China
E-mail: www.464234870@qq.com

Jianhui Pei
School of Computer Science and Engineering
ChongQing University of Technology
400054 ChongQing China
E-mail: 794349116@qq.com

Abstract—**The date mining based on big data was a very important field. In order to improve the mining efficiency, the mining algorithm of frequent itemsets based on mapreduce and FP-tree was proposed, namely, MAFIM algorithm. Firstly, the data were distributed by mapreduce. Secondly, local frequent itemsets were computed by FP-tree. Thirdly, the mining results were combined by the center node. Finally, global frequent itemsets were got by mapreduce and the search strategy. Theoretical analysis and experimental results suggest that MAFIM algorithm is fast and effective.**

Keywords-FP-tree; Mapreduce; Frequent itemsets; Big data; Data mining

I. INTRODUCTION

Data mining[1] was used to find a novel, effective, useful and understandable knowledge from the dataset. The main research directions of data mining include association rules[1] , classification, clustering and so on. The date mining based on big data[2,3] was a very important field. The key step of association rules was to get frequent itemsets[4,5] from dataset, and all frequent itemsets were subsets of maximal frequent itemsets. Therefore, all frequent itemsets could be found by mining maximal frequent itemsets. In order to improve the mining efficiency[6,7], the mining algorithm of frequent itemsets based on mapreduce and FP-tree was proposed, namely, MAFIM algorithm.

II. RELATED DESCRIPTION

A. Description of Mining Global Frequent Itemsets

The global transaction database[8,9] as DB, number of transaction as D. P_1、P_2、…、P_n, as the computer node, DB_i(i=1,2,…,n) as local transaction database for DB stored in the P_i node, the number of transaction is D_i, then $DB=\bigcup_{i=1}^{n} DB_i$, $D=\sum_{i=1}^{n} D_i$. Global frequent itemsets mining is through many nodes cooperation and finally dig out the global frequent item[10,11] E_{DB} of DB and the maximum frequent itemsets $F_{DB.}$

B. Description of Global Maximum Frequent Itemsets

Global transaction database as db, number of transaction as d.db_i (i=1,2,…,n) as local transaction database for db stored in the P_i node, the number of transaction is d_i , then $db=\bigcup_{i=1}^{n} db_i$, $d=\sum_{i=1}^{n} d_i$. Global maximal frequent itemsets used E_{DB} and F_{DB} which have been mined, and digging out the whole transaction database's global frequent item $DB \cup db$ and global maximal frequent itemsets $F_{DB \cup db}$.

C. Relevant Definition

Definition 1: To a set X, Local database $DB_i(i=1,2,\dots,n)$ includes X's transaction number, called local frequency of X in DB_i, use $X.\mathrm{si}_{\mathrm{DB}}$ as the symbol. The local frequency of X in db_i was $X.si_{db.}$

Definition 2: To a set X, Global transaction database DB includes X's transaction number, called global frequency of X in DB, use $X.s_{DB}$ as the symbol. The global frequency of X in db was $X.s_{db.}$

Definition 3: To a set X, if $X.si_{DB} \geq minsup \times D_i(i=1,2,\dots,n)$, called X is a local frequent itemsets of DB_i, all local frequent itemsets compose to F_{DB_i} , where minsup is the minimum support threshold. All local frequent itemsets in db_i compose to $F_{db\_i.}$

Definition 4: To a set X, if $X.s_{DB} \geq minsup \times D$, called X is a global frequent itemsets of DB, all global frequent itemsets compose to F_{DB} . All global frequent itemsets in db compose to $F_{db.}$

Definition 5: To sets X and Y, if $X \subseteq Y$, called X is a subset of Y, Y is a superset of X.

Definition 6: DB's global frequent itemsets X, if X is not a superset of all global frequent itemsets, called X is a global frequent itemsets of DB, all global frequent itemsets compose to F_{DB}. All global frequent itemsets in db compose to $F_{DB.}$

Definition 7: x_i is a item of DB, set X={ x_i }, if $X.s_{DB} \geq minsup \times D$, called x_i is a global frequent item of DB, all global frequent itemsets compose to $E_{DB.}$ All global frequent itemsets in db compose to $E_{db.}$

D. Relevant Theorem

Theorem 1: If the itemsets X is a global frequent itemsets for DB, then X is a local frequent itemsets in a local database DB $_i$ (i=1,2,…,n).

Prove: X is a global frequent itemsets for DB,satisfy $X.s_{DB} \geq \left(D_1 + D_2 + ... + D_n\right) \times \min\sup$. According to the Pigeonhole principle, there is at least one local database DB$_{i,}$ make $X.si_{DB} \geq min\,sup \times D_i$,so X is a local frequent itemsets of DB$_i$ theorem 1 established.

Theorem 2: If the itemsets X is a global maximum frequent itemsets for DB, then X is a subset of a local maximal frequent itemsets in a local database DB $_i$ (i=1,2,…,n).

Prove: itemset X is the global maximum frequent itemsets of DB. The itemset X is global frequent itemsets . According to theorem 1, X is a local frequent itemsets of DB$_i$ for a local database. According to the definition of 6, X is a subset of a local maximal frequent itemsets on the DB $_i$, theorem 2 established.

Theorem 3: The global maximum frequent itemsets of global transaction database DB and global increment transaction database A are respectively DB and B

The global maximum frequent itemsets of global transaction database DB and global increment transaction database db are F_{DB} and F_{db} respectively, the global maximum frequent itemset of DB ∪db is $F_{DB \cup db}$, for any set of $X \in F_{DB \cup db}$, both have itemset $Y \in F_{DB} \cup F_{db}$, promote $X \subseteq Y$.

Prove: If itemset X is any of any global maximum frequent itemsets, according to theorem 2, X may be a subset of the global maximal frequent itemsets in DB, and may be a subset of the global maximal frequent itemsets in db, theorem 3 established.

Theorem 4: E$_{DB}$ is the global frequent item of DB which according to the support component in descending order, E$_{db}$ is the global frequent item of db which according to the support component in descending order, all items in E$_{DB}$∩E$_{db}$ are global frequent items in DB∪db.

Prove: If item X is any one of E$_{DB}$∩E$_{db,}$ then x is not only a global frequent items of DB, but also a global frequent items of db, X={x}, that X.s$_{DB}$≥minsup×D and X.s$_{db}$≥minsup×d,therefore,X.s$_{DBUdb}$=X.s$_{DB}$+X.s$_{db}$≥minsup×(D +d), theorem 4 established.

III. MAFIM ALGORITHM

MAFIM algorithm was proposed. Firstly, the data were distributed by mapreduce. Secondly, local frequent itemsets were computed by FP-tree. Thirdly, the mining results were combined by the center node. Finally, global frequent itemsets were got by mapreduce and the search strategy.

The pseudocode of MAFIM is described as follows.

Alogrithm MAFIM

Input: The local transaction database DB_i which has M_i tuples and $M = \sum_{i=1}^{n} M_i$, n nodes P_i(i=1,2,…n), the center node P_0, the minimum support threshold *min_sup*.

Output: The global frequent itemsets F.

Methods: According to the following steps.

```
step1. /* the data were distributed by mapreduce*/
        for($i$=1;$i$<=$n$;$i$++)
            $P_0$ transmits $DB_i$ to $P_i$ ;
Step2. /*local frequent itemsets were computed by FP-tree*/
        for($i$=1;$i$<=$n$;$i$++)
        {   creating the $FP\text{-}tree^i$;
            $F_i$ =FP-growth($FP\text{-}tree^i$, null);
        }
step3./* the mining results were combined by the center node*/
        for($i$=1;$i$<=$n$;$i$++)
            $P_i$ sends $F_i$ to $P_0$;
        $P_0$ combines $F_i$ and produces $F'$ ; /* $F' = \bigcup_{i=1}^{n} F_i$ */
Step4./*global frequency of itemsets were computed*/
        for each items $d \in$ the remain of $F'$
            $d.s = \sum_{i=1}^{n} d.si$ ;
step5./*global frequent itemsets were got by mapreduce and the search strategy*/
        for each items $d \in$ the remain of $F'$
            if ($d.s$>=$min\_sup$*$M$)
                $F$=$F \cup d$;
```

IV. EXAMPLE OF MAFIM ALGORITHM

With 3 stations P1, P2 and P3, corresponding to a local database DB1, DB2 and DB3. Each database as shown in table I. Minimum support threshold min_sup=0.42.

TABLE I. LOCAL DATABASE DB1, DB2, DB3

Local database	**ID**	**Transaction**
DB1	100	a，b，c，k，m，f，e，l，p
	101	c，k，b，m，o，q
	102	a，b，c，d，e
DB2	200	f，h，j，q
	201	a，b，c，m，l，f，k
	202	c，r，s，t，q
DB3	300	a，b，c，d，e，f
	301	b，c，d，k，q
	302	f，s，m，q

According to table 1 and min_sup=0.42, can draw the global frequent items, in accordance with the degree of support in descending order, as shown in Table II.

TABLE II. GLOBAL FREQUENT ITEMS AND SUPPORT COUNT

Global frequent Items	Support count(Global frequency)
c	7
b	6
f	5
q	5
a	4
m	4
k	4

The global frequent itemset composed of E={c, b, f, q, a, m, k}.

The search strategy implementation process as shown in table III. Min_sup=0.42.

TABLE III. THE SEARCH STRATEGY

F'	Set the length of K	The search strategy	F
{{ c, b, m, k}, {c, b, a}, {c, b}}	*4*	*{ c, b, m, k}*	
{{c, b, a}, {c, b, m}, {c, b, k}, {b, m, k}, {c, b}}	*3*	*{c, b, a} ✓*	*{{c, b, a}}*
{{c, b, m}, {c, b, k}, {b, m, k}}	*3*	*{c, b, m}*	*{{c, b, a}}*
{{c, b, k}, {b, m, k}, {b, m}, {c, m}}	*3*	*{c, b, k} ✓*	*{{c, b, a}, {c, b, k}}*
{{b, m, k}, {b, m}, {c, m}}	*3*	*{b, m, k}*	*{{c, b, a}, {c, b, k}}*
{{b, m}, {c, m}, { m, k}}	*2*	*{b, m}*	*{{c, b, a}, {c, b, k}}*
{{c, m}, { m, k}}	*2*	*{c, m}*	*{{c, b, a}, {c, b, k}}*
{{ m, k}}	*2*	*{ m, k}*	*{{c, b, a}, {c, b, k}}*

V. EXPERIMENTS OF MAFIM

MAFIM compares with CD and FDM in terms of communication traffic and runtime. The experimental data comes from the sales data in July 2015 of a supermarket. The results are reported in Fig .1 and Fig .2.

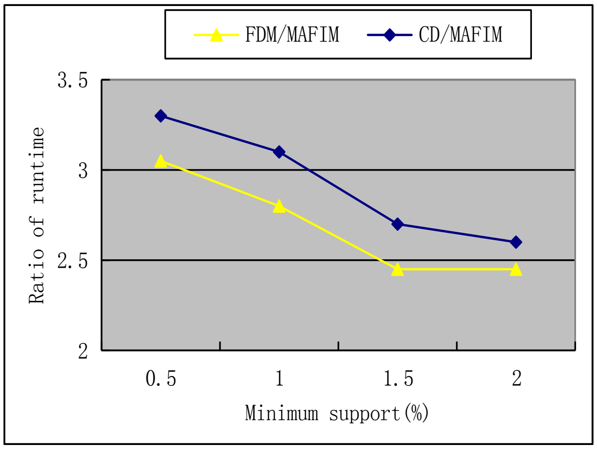


Figure 1. Comparison of communication traffic

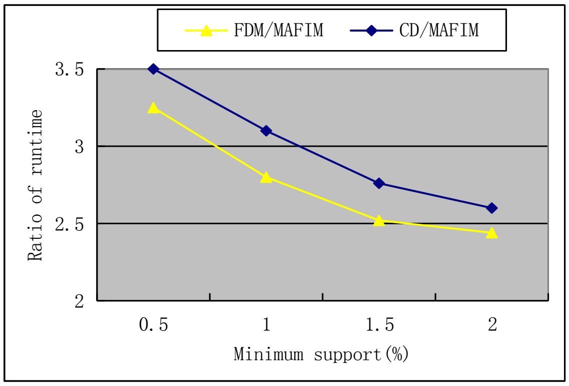


Figure 2. Comparison of runtime

The comparison experiment results indicate that under the same minimum support threshold, the communication traffic and runtime of MAFIM decreases while comparing with CD and FDM.

VI. CONCLUSION

The mining algorithm of frequent itemsets based on mapreduce and FP-tree was proposed. Firstly，the data were distributed by mapreduce. Secondly, local frequent itemsets were computed by FP-tree. Thirdly, the mining results were combined by the center node. Finally, global frequent itemsets were got by mapreduce. It can promote highly the efficiency of data mining.

ACKNOWLEDGMENT

This research is supported by scientific and technological research program of chongqing municipal education commission under grant No.KJ1709214. This research is supported by the national statistical science research project under grant No.2015LZ22 and the humanities and social science research project of chongqing municipal education commission under grant No.15SKG131.

REFERENCES

[1] Han JW, Kamber M, Pei J. Data Mining: Concepts and Techniques Third Edition [M]. San Francisco: Morgan Kaufmann, 2011.

[2] Big Data Across the Federal Government [EB/OL]. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf, 2012.

[3] Science. Special Online Collection: Dealing with Data [EB].

http://www.sciencemag.org/site/special/data/, 2011.

[4] Marconi K, Lehmann H. Big Data and Health Analytics[M]. Boca Raton:CRC Press, 2014.

[5] He B, Yan H. Incremental Updating Algorithm of Global Maximum Frequent Itemsets in Distributed Database[J]. Journal of Sichuan University(Engineering Science Edition), 2012,44(3):112~117. (in Chinese with English abstract)

[6] McKinsey&Company. The big-data revolution in US health care: Accelerating value and innovation [R]. http://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care, 2013.

[7] He B. Fast Mining of Global Maximum Frequent Itemsets in Distributed Database [J]. Control and Decision, 2011,26(8):1214~1218. (in Chinese with English abstract)

[8] Muin J. Khoury and John P. A. Ioannidis. Big data meets public health[J]. Science, 2014, 346(6213) : 1054-1055.

[9] Chen ZB, Han H, Wang JX. Data Warehouse and Data Mining[M].Beijing: Tsinghua University Press, 2009.

[10] Song YQ, Zhu ZH, Chen G. An algorithm and its updating algorithm based on FP-tree for mining maximum frequent itemsets[J]. Journal of software, 2003,14(9):1586~1592(in Chinese with English abstract)

[11] Bayardo RJ. Efficiently mining long patterns form databases[C]. In: Haas LM, Tiwary A, eds. Proc. Of the ACM SIGMOD International Conference on Management of Data. Dallas:ACM Press, 2000. 1~12.

2017 International Conference on Computer Network, Electronic and Automation (ICCNEA 2017)

Self-adaptive Differential Evolutionary Extreme Learning Machine and Its Application in Facial Age Estimation

Junhua Ku
Department of information engineering
Hainan institute of science & technology Haikou, China
E-mail: kujunhua@163.com

Kongduo Xing
Department of information engineering
Hainan institute of science & technology Haikou, China
E-mail: cogemm@163.com

Abstract—In this paper, Self-adaptive Differential Evolutionary Extreme Learning Machine (SaDE-ELM) was proposed as a new class of learning algorithm for single-hidden layer feed forward neural network (SLFN). In order to achieve good generalization performance, SaDE-ELM calculates the error on a subset of testing data for parameter optimization. Since SaDE-ELM employs extra data for validation to avoid the over fitting problem, more samples are needed for model training. In this paper, the cross-validation strategy is proposed to be embedded into the training phase so as to solve the overtraining problem. Experimental results demonstrate that the proposed algorithms are efficient for Facial Age Estimation.

Keywords-Extreme learning machines; Differential evolution extreme learning machines; Self-adaptive differential evolution extreme learning machines; Facial age estimation

I. INTRODUCTION

Automated age estimation from facial images is one of the most difficult challenges in face analysis [1,2]. It can be very favorable in a number of real life applications such as age-based authorization systems, demographic data mining, business intelligence and video surveillance systems. The difficulty of this task originates from many reasons such as the lack of enough labeled samples to model the aging patterns of subjects, as well as uncontrolled conditions in data collection such as illumination, pose, occlusions and other environmental variables. Aging process is also known to be very subject-dependent, i.e. subjects might differ in terms of aging patterns, resulting in high variations within the samples from the same age.

Recently, a new fast learning neural algorithm for SLFNs, named extreme learning machine (ELM) [3,4], was developed to improve the efficiency of SLFNs. Different from the conventional learning algorithms for neural networks (such as BP algorithms[5]), which may face difficulties in manually tuning control parameters (learning rate, learn-ing epochs, etc.) and/or local minima, ELM is fully auto-matically implemented without iterative tuning, and in theory, no intervention is required from users. Furthermore, It was popular for its fast training speed by means of utilizing random hidden node parameters and calculating the output weights with least square algorithm [6-10]. However, in ELM, the number of hidden nodes is assigned a priori, the hidden node parameters are randomly chosen and they remain unchanged during the training phase. Many non-optimal nodes may exist and contribute less in minimizing the cost function. Moreover, in [11] Huang et al. pointed out that ELM tends to require more hidden nodes than conventional tuning-based algorithms [12, 13] in many cases.

Differential evolution (DE) [14] which is a simple but powerful population-based stochastic direct searching technique is a frequently used method for selecting the network parameters [15–17]. In [15], DE is directly adopt-ed as a training algorithm for feed forward networks where all the network parameters are encoded into one population vector and the error function between the network approximate output and the expected output is used as the fitness function to evaluate all the populations. However, Subudhi and Jena [16] have pointed out that using the DE approach alone for the network training may yield a slow convergence speed. Therefore, in [17], a new algorithm named evolutionary extreme learning machine (DE-ELM) based on DE and ELM has been developed for SLFNs. Using the DE method to optimize the network input parameters and the ELM algorithm to calculate the network output weights, DE-ELM has shown several promising features. It not only ensures a more compact network size than ELM, but also has better generalization performance.

However, in the above DE based neural network training algorithms, the trial vector generation strategies and the control parameters in DE have to be manually chosen. For example, the control parameters in DE-ELM are manually selected according to an empirical suggestion and the simple random generation method is adopted to produce the trial vector. As pointed out by many researchers, the performance of the DE algorithm highly depends on the chosen trial vector generation strategy and the control parameters, and inappropriate choices of strategies and control parameters may result in premature convergence or stagnation. Therefore, we propose a novel learning algorithm named self-adaptive evolutionary extreme learning machine (SaDE-ELM) for SLFNs. In SaDE-ELM, the hidden node learning parameters are optimized by the self-adaptive differential evolution algorithm.

The rest of the paper is organized as follows. In section II, a brief introduction to ELM and SaDE are given. In Section III, we introduce model of proposed SaDE-ELM algorithm in detail. In Section IV, we present Performance Evaluation. In section V, we conclude and summarize our results.

II. BACKGROUND

As a novel training algorithm for SLFNs, ELM is very efficient and effective. In this section, we will give a brief review of ELM. In this section, we briefly review ELM and SaDE-ELM approach for ELM is the foundation of SaDE-ELM.

A. *Extreme Learning Machine (ELM)*

For *N* arbitrary distinct samples $(\mathrm{x}_j, \mathrm{t}_j)$, where $\mathrm{x}_j = [x_{j1}, x_{j2}, \cdots, x_{jn}]^T \in \mathbb{R}^n$, $\mathrm{t}_j = [t_{j1}, t_{j2}, \cdots, t_{jm}]^T \in \mathbb{R}^m$, SLFNs with *L* hidden nodes and activation function *g*(*x*) are

$$\sum_{i=1}^{L} \beta_i \, \mathrm{g}_i(\mathrm{x}_j) = \sum_{i=1}^{L} \beta_i \, \mathrm{g}_i(\mathrm{w}_i \cdot \mathrm{x}_j + b_j) = \mathrm{o}_j \qquad (j = 1, 2, ..., N) \qquad (1)$$

where **w***i*= [*w*i1 , *w*i2 , ..., *win*]*T* is the weight vector connecting the *i*th hidden node and the input nodes, β_i = [βi1 , βi2, ..., β_{im}]^T is the weight vector connecting the *i*th hidden node and the output nodes, b_i is the threshold of the *i*th hidden node, $\mathrm{W}_i \cdot \mathrm{X}_j$ denotes the inner product of w$_i$ and x$_j$, *g*(*x*) is activation function and Sigmoid, Sine, Hardlim and other functions are commonly used. The output nodes are chosen linear in this paper, and $\mathrm{o}_j = [o_{j1}, o_{j2}, ..., o_{jm}]^T$ is the *j*th output vector of the SLFNs [22].
The SLFNs with *L* hidden nodes and activation function g(x) can approximate these N samples with zero error. It means $\sum_{j=1}^{L} \left\| \mathrm{o}_j - \mathrm{t}_j \right\| = 0$ and there exist β_i , $\mathbf{w}_i$ and b_i such that

$$\sum_{i=1}^{L} \beta_i \, \mathrm{g}_i(\mathrm{x}_j) = \sum_{i=1}^{L} \beta_i \, \mathrm{g}_i(\mathrm{w}_i \cdot \mathrm{x}_j + b_j) = \mathrm{t}_j \qquad (j = 1, 2, ..., N) \qquad (2)$$

The equation above can be expressed compactly as follows:

$$\mathbf{H} \circledR = \mathbf{T} \qquad (3)$$

Where $\mathrm{H}(\mathrm{w}_1, \mathrm{w}_2, \cdots, \mathrm{w}_L, b_1, b_2, \cdots, b_L, \mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_L)$

$$=[h_{ij}]=\begin{bmatrix} \mathrm{g}(\mathrm{w}_1 \cdot \mathrm{x}_1 + b_1) & \mathrm{g}(\mathrm{w}_1 \cdot \mathrm{x}_1 + b_2) & \cdots & \mathrm{g}(\mathrm{w}_1 \cdot \mathrm{x}_1 + b_L) \\ \mathrm{g}(\mathrm{w}_1 \cdot \mathrm{x}_2 + b_1) & \mathrm{g}(\mathrm{w}_2 \cdot \mathrm{x}_2 + b_2) & \cdots & \mathrm{g}(\mathrm{w}_L \cdot \mathrm{x}_2 + b_L) \\ \vdots & \vdots & & \vdots \\ \mathrm{g}(\mathrm{w}_1 \cdot \mathrm{x}_N + b_1) & \mathrm{g}(\mathrm{w}_2 \cdot \mathrm{x}_N + b_2) & \cdots & \mathrm{g}(\mathrm{w}_L \cdot \mathrm{x}_N + b_L) \end{bmatrix}_{N \times L}$$

$$\beta = \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{L1} & \beta_{L2} & \cdots & \beta_{Lm} \end{pmatrix} \text{ and } \quad \mathrm{T} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ t_{N1} & t_{N2} & \cdots & t_{Nm} \end{pmatrix}$$

The matric **H** is called the hidden layer output matrix of the neural network and the *i*th column of **H** is the *i*th hidden node output with respect to inputs **x**$_1$, **x**$_2$, ..., **x**$_N$.

By simply randomly choosing hidden nodes and then adjusting the output weights, single hidden layer feedforward networks (SLFNs) work as universal approximators with any bounded non-linear piecewise continuous functions for additive nodes [23]. ELM algorithm claims that the hidden node parameters can be randomly assigned [3,4], then the system equation becomes a linear model and the network output weights can be analytically determined by finding a least-square solution of this linear system as follow

$$\widehat{\beta} = \mathbf{H}^{\dagger}\mathbf{T} \qquad (4)$$

Where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse of matrix **H**. Then the output function of ELM can be modeled as follows.

$$\phi(\xi) = \gamma(\xi) \circledR = \gamma(\xi)\mathbf{H}^{\square}\mathbf{T} \qquad (5)$$

Moreover, it should be noted that many nonlinear activation and kernel functions can be used in ELM.

B. *Self-adaptive Differential Evolution*

Differential evolution (DE), proposed by Storn and Price in 1995, is a simple yet powerful evolutionary algorithm (EA) [24]. There are three parameters in DE algorithm, which are the population size *NP*, mutation scaling factor *F* and crossover rate *CR*. *NP* is a problem-dependent parameter, while *F* and *CR* are very sensitive to the performance at different stages of evolution. To overcome the limitations of choosing the parameters in DE, Brest et al. proposed a parameter adaptation technique to choose the mutation scaling factor *F* and crossover rate *CR* namely SADE-ELM algorithm which performs better than the basic DE algorithm. In general, SADE algorithm is composed of three main steps: mutation, crossover, and selection [26].
We consider the following optimization problem:

Minimize $f(x_i)$, $x_i \in R_D$

where $x_i = [x_{i1}, x_{i2}, \cdots, x_{iD}]^T, i = 1, 2, \cdots, NP$ is a target vector of D decision variables. During the mutation operation, mutant vector v$_i$ is generated by mutation strategy in the current population:

$$v_i = x_{r1} + F \cdot (x_{r2} - x_{r3}) \qquad (6)$$

where $r1, r2, r3$ are mutually exclusive integers randomly chosen in the range [1,*NP*], and $r1 \neq r2 \neq r3 \neq i$.
Following mutation, trial vector u_i is generated between x_i and v_i during crossover operation where the most widely used operator is the binomial crossover performed as follows:

$$u_{ij} = \begin{cases} v_{ij}, & \textit{if}\,(\mathrm{rndreal}(0,1) < \mathrm{CR} \;\text{ or }\; j = j_{\mathrm{rand}}), \\ x_{ij}, & \textit{otherwise} \end{cases} \qquad (7)$$

Where j_{rand} is a integer randomly chosen in the range [1,*D*], and rndreal(0, 1) is a real number randomly generated in (0, 1). Finally, to keep the population size constant during the evolution, the selection operation is used to determine whether the trial or the target vector survives to the next generation according to one-to-one selection:

$$x_i = \begin{cases} u_{i,} & if\,(f(u_i)\,f(x_i)) \\ x_i, & otherwise \end{cases} \tag{8}$$

Where *f*(*x*) is the optimized objective function. During the evolution, *F* and *CR* are adaptively tuned to improve the performance of DE for each individual

$$F_{i,G+1} = \begin{cases} F_l + rand_1 \cdot F_u & if\,(rand_2 < \tau_1) \\ F_{i,G} & \text{otherwise} \end{cases} \tag{9}$$

$$CR_{i,G+1} = \begin{cases} rand_3 & if\,(rand_4 < \tau_2) \\ CR_{i,G} & \text{otherwise} \end{cases} \tag{10}$$

Where $F_{i;G+1}$ and $CR_{i;G+1}$ are the mutation scaling factor and crossover rate for *i* individual in *G* generation respectively, *randj*=1;2;3;4 are randomly chosen from (0, 1), τ_1 and τ_2 both valued 0.1 which is used to control the generation of *F* and *CR, Fl* valued 0.1 and *Fu* is valued 0.9. In the first generation, *F* and *CR* are initialized to 0.5.

C. Model of Proposed SADE-ELM Algorithm

Since the ELM generates the input weights and hidden biases arbitrarily which are the basic of calculating the output weights, it may not reach the optimal result in classification or regression. Thus, a hybrid approach integrated self-adaptive differential evolution algorithm and extreme learning machine namely SADE-ELM algorithm to optimize the input weights and hidden biases is able to obtain better generalization performance than ELM algorithm [17].

In SaDE-ELM, we proposed SaDE-ELM for SLFNs by incorporating the self-adaptive differential evolution algorithm [25] to optimize the network input weights and hidden node biases and the extreme learning machine to derive the network output weights.

Given a set of training data and L hidden nodes with an activation function g(·), we summarize the SaDE-ELM algorithm in the following steps.

Step 1. Initialization

A set of *NP* vectors where each one includes all the network hidden node parameters are initialized as the populations of the first generation

$$\theta_{k,G} = \left[\mathrm{w}_{1,k,G}^T, \cdots, \mathrm{w}_{L,k,G}^T, b_{1,k,G}^T, \cdots, b_{L,k,G}^T \right] \tag{11}$$

where w_j and b_j (*j* = 1, … , *L*) are randomly generated, *G* represents the generation and *k* = 1, 2,…, *NP*.

Step 2. Calculations of output weights and RMSE

Calculate the network output weight matrix and root mean square error (RMSE) with respect to each population vector with the following equations, respectively.

$$\beta_{k,G} = \mathrm{H}_{k,G}^{\dagger}\mathrm{T} \tag{12}$$

$$\mathrm{RMSE}_{k,G} = \sqrt{\frac{\sum_{i=1}^{N} \left\| \sum_{j=1}^{L} \beta_j g(\mathrm{w}_{j,k,G}, b_{j,k,G}, x_i) - t_i \right\|}{m \times N}} \tag{13}$$

Then use the value of RMSE to calculate the new best population vector $\theta_{k,G+1}$ with the following equation.

$$\theta_{k,G+1} = \begin{cases} u_{k,G+1} & if\ (\mathrm{RMSE}_{\theta_{k,G}} \text{-} \mathrm{RMSE}_{u_{k,G+1}}) > \varepsilon \cdot \mathrm{RMSE}_{\theta_{k,G}} \\ u_{k,G+1} & if\ \left| \mathrm{RMSE}_{\theta_{k,G}} \text{-} \mathrm{RMSE}_{u_{k,G+1}} \right| < \varepsilon \cdot \mathrm{RMSE}_{\theta_{k,G}} \\ & \quad and \left\| \beta_{u_{k,G+1}} \right\| < \left\| \beta_{\theta,G} \right\| \\ \theta_{i,G} & \text{otherwise} \end{cases} \tag{14}$$

where ε is the preset small positive tolerance rate. In the first generation, the population vector with the best RMSE is stored as $\theta_{\mathrm{best},1}$ and RMSE$\theta_{\mathrm{best},1}$.

All the trial vectors $u_{k,G+1}$ generated at the (G+1)th generation are evaluated using equation(11) .The norm of the output weight $\|\beta\|$ is added as one more criteria for the trial vector selection as pointed out in [13] that the neural networks tend to have better generalization performance with smaller weights.

The three operations mutation, crossover and selection are repeated until the preset goal is met or the maximum learning iterations are completed. At last we calculate the output weigh $\beta = [\beta_{i1}\ \beta_{i2}\ \cdots\ \beta_{iL}]^T$ with equation $\beta = \mathbf{H}^{\dagger}\mathbf{T}$.

III. Performance Evaluation

In this section we describe the different parts of our age estimation pipeline, namely face alignment, feature extraction and model learning. The workflow of our proposed method is illustrated in Fig .1.

The first steps of a human age estimation pipeline are face detection [17, 18] and facial landmark localization]. In this work, we chose to use the Deformable Part Model (DPM) based face detector proposed by Mathias et al. [18], because it finds the location of the face bounding box and gives a good alignment without the need for facial landmark localization. The DPM face detector gives the coordinates of the bounding box (if any face is detected), as well as the detection score. We run the face detector on rotated version of the original image between -60◦ and 60◦ in 5◦ increments, in order to eliminate in-plane rotation. Since some of the images are rotated 90° or upside down, we also try 180°, -90° and 90° rotations. We then take the output with the maximum face score. For the cases where no face is detected, we register the whole image. ChaLearn Looking at People 2016 - Apparent Age Estimation challenge dataset [19] consists of 7,591 face images collectively labeled by multiple human annotators, therefore the mean μ and the standard deviation σ is provided for each sample. The dataset is split into 4113 training, 1500 validation and 1978 testing samples, where the testing set labels are sequestered. The three subsets have a similar age distribution. Table I presents the number of samples where the DPM face detector was able to detect a face. Table I shows the number of detections on the three subsets.

TABLE I. FACE ALIGNMENT SUMMARY

#	Train	Val	Test
Given	4113	1500	1978
Detected	4016	1462	1920

We used the deep network to extract CNN features from aligned images. The VGG- Face network consists of 37 layers, the final one being a 2622-dimensional softmax layer, trained for the face recognition task. We tried the performance of the final layers and found that the 33rd layer, which is the first (earliest) 4096-dimensional convolution layer, was the most informative one. Therefore we used only the features from this layer in model learning. The baseline regression performances (without any grouping) of the best layers are shown in Table II

TABLE II. COMPARISON OF DIFFERENT LAYERS OF VGG-FACE

Layer	Num. features	val	MAE$_{\text{val}}$
32	25088	0.4284	4.68
33	4096	0.4021	4.35
35	4096	0.4150	4.48
37	2622	0.4066	4.38

We then normalize each feature vector by dividing it to its Euclidean norm. We have tried various normalization options prior to L2 normalization and saw that none of them was improving the normal score; therefore we decided to use only L2 normalization for the final system. Performance with various normalization options for the best layers is shown in Table III.

In our experiments, we tried combinations of alternative feature normalization methods, including the sigmoid function, power normalization by 2 (i.e. setting the absolute value of each feature to its square root), min-max normalization of each feature to [−1, 1] among samples, and z- normalization. For min-max and z-normalization, we learn the parameters from training folds and apply them to the test fold.

TABLE III. Validation set performance with different normalization options

Norm. Type	Lay r 33		Lay r 37	
	ε	MAE	ε	MAE
Nonorm	0.4487	4.91	0.4403	4.79
L2	0.4021	4.35	0.4066	4.38
Pow. + L2	0.4028	4.32	0.4079	4.44
Sig. + L2	0.4152	4.49	0.4137	4.46
MM+L2	0.4355	4.77	0.4301	4.63
Z+L2	0.4102	4.48	0.4036	4.33
MM +Sig. +	0.4861	5.46	0.4652	5.12
Z + Sig. + L2	0.4220	4.59	0.4164	4.51
M 1 + Pow. + L2	0.4565	5.01	0.4438	4.88
Z + Pow. + L2	0.4083	4.43	0.4078	4.37

Now, we introduce the evaluation criteria in our experiments as following.

Mean Absolute Error (MAE): A standard way of measuring the accuracy of a regressor is to average the absolute deviation of each sample’s label from its estimated value. More formally, MAE of a given dataset is calculated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | x_i - \hat{x}_i | \tag{15}$$

where x_i is the true label i.e. the average of apparent age annotations for sample *i*, $\hat{x}_i$ is the predicted value, and *N* is the number of testing samples.

Normal Score (ε): Since the LAP-2016 dataset is labeled by multiple annotators, the performance of an age estimation system might be more accurately measured by taking into account the variance of the annotations for each sample. Therefore the ϙ-score is calculated by fitting a normal distribution with mean μ and standard deviation σ of the annotations for each sample:

$$\varepsilon = 1 - \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \tag{16}$$

Thus, the average ϙ-score for a dataset can change between 1 (worst case) and 0 (best case).

The system is implemented in MATLAB. Face detection takes around 2 seconds per image and rotation angle. Feature extraction from VGG-Face with MatConvNet library takes around 1 second per image. For classification and regression, we optimize the kernel parameter γ and the regularization coefficient C with a grid search where both parameters are searched in the exponential set $2^{[-2,-1, \ldots ,6]}$. Training the whole system takes 12 minutes and obtaining the estimation takes around 2 seconds per test image.

According to the above conditions, we present the results of our classification and regression systems. In Table 4, we summarize the classification accuracy and recall for the 8 overlapping age groups we used. The 9th row is the performance of the backup system, and the final row is the performance of the whole system on the validation set of LAP-2016 dataset.

Table Ⅳ shows that the ensemble of local regressors yield smaller MAE for younger age groups. As the age progresses, within-group variance increases with it, making the apparent age estimation task harder. Finally, since younger subjects are usually annotated with less variance, the ϙ-score behaves almost inversely to MAE score, as the ϙ-score is more tolerant for the errors in the older subjects. We display the estimation results on samples from the validation set in Fig. 2, which shows the invariance of CNN features to common difficulties such as blur, pose and occlusions.

TABLE IV. CLASSIFICATION ACCURACY, RECALL AND REGRESSION PERFORMANCE ON VALIDATION SET WITH DIFFERENT AGE GROUPS. N DENOTES THE NUMBER OF SAMPLES

Group	**N_{tr}**	**N_{val}**	**Acc.**	**Rec.**	ε	**MAE**
0-15	860	152	0.96	0.78	0.45	2.46
10-25	2366	436	0.84	0.65	0.31	2.90
15-30	3686	662	0.84	0.83	0.31	3.19
20-35	4072	705	0.81	0.86	0.33	3.52
30-40	1764	311	0.81	0.35	0.34	3.82
35-50	1568	288	0.85	0.45	0.34	4.26
45-60	976	184	0.91	0.48	0.28	3.87
55-∞	554	106	0.96	0.57	0.28	4.36
0-∞	8032	1462	-	-	0.40	4.35
Overall	8032	1462	-	-	0.33	3.85

IV. CONCLUSION

In this paper, we propose an apparent age estimation system with the use of SaDE-ELM Algorithm. We show that the performance of local regressors are better than the global regressor for almost all groups. However, we give equal weight to each group a sample is assigned to, whereas weighing the decisions with a membership score can result in more accurate estimation. CNNs are robust to common difficulties in image processing such as pose and illumination differences as well as occlusions. Therefore our system works with a very coarse alignment system, however we believe that obtaining a finer alignment with the help of a landmark detection system will further improve the estimation accuracy.We make use of transfer learning by using the features from a deep network that is trained on a face recognition task and directly employing them in age estimation.

ACKNOWLEDGMENT

This research was supported by science research project of Education Department of Hainan province (Hnky2017ZD-20).

REFERENCES

[1] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. IEEE Trans. Syst. Man Cybern. B, 34(1):621–628, 2004.

[2] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. TPAMI, 24(4):442–455, 2002.

[3] Huang GB, Zhu QY, Siew CK. Extreme learning machine: a new learning scheme of feedforward neural networks. In: Proceedings of international joint conference on neural networks (IJCNN2004), vol 2, no 25–29, pp 985–990.

[4] Huang GB, Zhu QY, Siew CK. Extreme learning machine: theory and applications. Neurocomputing 70(1–3):489–501.

[5] Espana-Boquera S, Zamora-Mart ńez F, Castro-Bleda M J, et al. Efficient BP algorithms for general feedforward neural networks[C]//International Work-Conference on the Interplay Between Natural and Artificial Computation. Springer Berlin Heidelberg, 2007: 327-336.

[6] G. Thatte, U. Mitra, and J. Heidemann, "Parametric methods for anomaly detection in aggregate traffic," IEEE/ACM Transactions on Networking, vol. 19, no. 2, pp. 512–525, April 2011.

[7] M. Qin and K. Hwang, "Frequent episode rules for internet anomaly detection," in Proceedings of the Network Computing and Applications, Third IEEE International Symposium. Washington, DC, USA: IEEE Computer Society, 2004, pp. 161–168.

[8] X. He, C. Papadopoulos, J. Heidemann, U. Mitra, and U. Riaz, "Remote detection of bottleneck links using spectral and statistical methods," Computer Networks, vol. 53, pp. 279–298, February 2009.

[9] W. W. Streilein, R. K. Cunningham, and S. E. Webster, "Improved detec- tion of low-profile probe and denial-of-service attacks," in Proceedings of the 2001 Workshop on Statistical and Machine Learning Techniques in Computer Intrusion Detection, June 2001.

[10] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273–297, 1995.

[11] G.-B. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: a survey," International Journal of Machine Leaning and Cybernetics, vol. 2, no. 2, pp. 107–122, 2011.

[12] G. Tandon, "Weighting versus pruning in rule validation for detecting network and host anomalies," in In Proceedings of the 13th ACM SIGKDD international. ACM Press, 2007.

[13] Y. Liao and V. R. Vemuri, "Use of k-nearest neighbor classifier for intrusion detection," Computers & Security, vol. 25, pp. 439–448, 2002.

[14] Storn R, Price K (2004) Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J Glob Optim 11(4):341–359

[15] Ilonen J, Kamarainen JI, Lampinen J (2003) Differential evolution training algorithm for feedforward neural networks. Neural Process Lett 17:93–105

[16] Subudhi B, Jena D (2008) Differential evolution and levenberg marquardt trained neural network scheme for nonlinear system identification. Neural Process Lett 27:285–296.

[17] P. Viola and M. J. Jones. Robust real-time face detection. IJCV, 57(2):137–154, 2004.

[18] [18] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In ECCV, 2014, pages 720–735.

[19] S. Escalera, M. Torres, B. Martnez, X. Bar, H. J. Escalante, I. Guyon, G. Tzimiropoulos, C. Corneanu, M. Oliu, M. A. Bagheri, and M. Valstar. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In ChaLearn Looking at People and Faces of the World, CVPRW, 2016

.

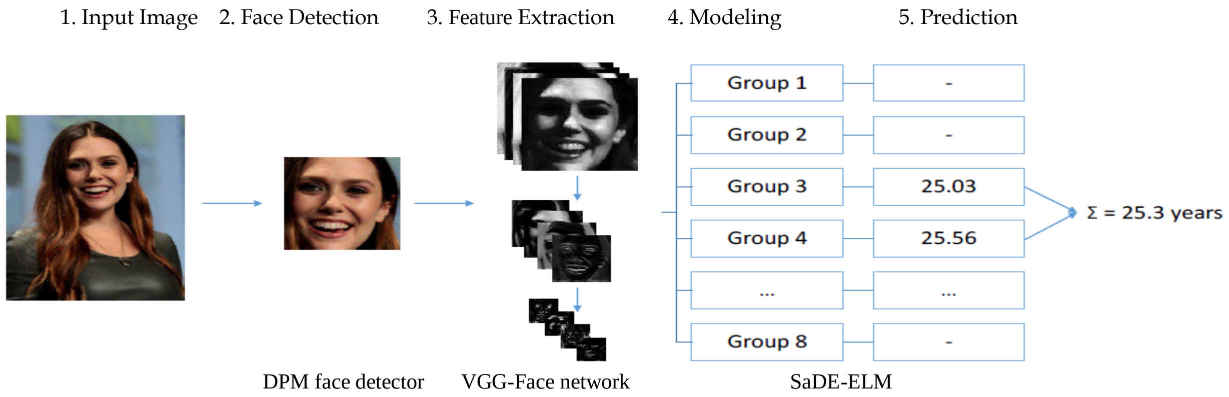
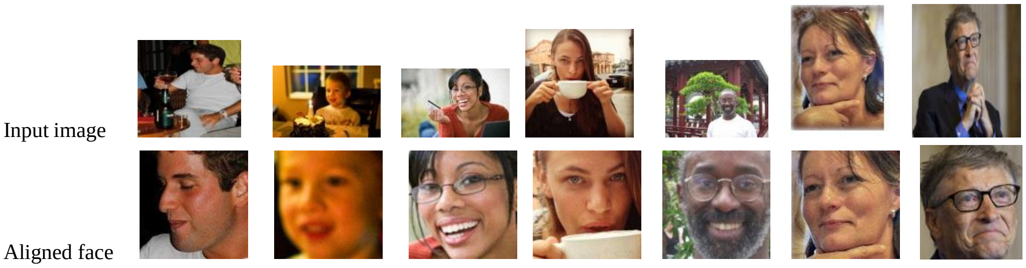


Figure 1. Pipeline of the proposed system



Apparent age	23	4	28	25	49	48	63
Predicted age	23.57	4.15	27.84	26.87	47.94	46.79	61.83

Figure 2. Application in facial age estimation based on sade-elm from the validation set

2017 International Conference on Computer Network, Electronic and Automation (ICCNEA 2017)

GrandStore: Towards Large-Scale Free Personal Cloud Storage

Li Zhang,
School of Computer Science and Engineering,
Hunan University of Science and Technology,
Xiangtan, Hunan, 411201, P.R. China
zlhncdsy@163.com

Bing Tang
School of Computer Science and Engineering,
Hunan University of Science and Technology,
Xiangtan, Hunan, 411201, P.R. China
zlhncdsy@163.com

Abstract—Personal cloud storage services are gaining popularity, such as SkyDrive, iCloud, Dropbox, etc. All of them provide a certain amount of free storage space for individual users, while the free space is quite limit, and you should upgrade to a paid account to get extra space. Therefore, a new approach is proposed in this paper, that many free personal cloud storage accounts are integrated in order to realize large-scale free personal cloud storage. A prototype system called GrandStore is designed and implemented, which is based on the principle of OAuth protocol and open API. Specifically, after authorized by the owner of account, GrandStore could manage and control the account, so there is no need for complex login any more. Users only need apply for several free cloud storage accounts, and then account authentication credentials are stored in back-end database of GrandStore, which realizes easily enlarging personal free storage space, and managing all storage space in a unique access entry.

Keywords-Cloud storage; GrandStore; OAuth 2.0; Open platform

I. INTRODUCTION

Data explosion is one of the biggest issues facing IT today. The amount of data that organizations store has grown exponentially in the last 10 years. How to store and manage these large-scale data is really a great problem. One solution to this problem is using cloud storage, an infrastructure that provides on-demand online storage services over Internet. Cloud storage is now the new direction of storage technology, which uses virtualized and scalable storage resource pool to provide storage service for users. It is allowed to use all kinds of method to consume cloud storage service through Internet, such as Web, client program and open interface, following the rule of pay-as-you-go. Cloud storage could deliver online services to individuals or companies, including online file hosting, storage and backup. Recently, personal cloud storage services are very popular and attract our attention, such as Microsoft SkyDrive, Apple iCloud, Google Drive, Dropbox, etc. All of them provide free storage space for individuals, as well as file synchronization service, while the free space is quite limit, and you should pay for extra free space.

Since that there are a variety of free personal cloud storage services, basically users should register accounts to use these services. Usually, one user has several accounts, and these accounts belong to different personal cloud storage providers. In this situation, we are confronted with two great problems. First, how to manage multi-accounts of different cloud storage providers using a unique access entry; second, how to obtain more free storage space, since the free space is quite limit.

To tackle these two challenges, in this paper we propose an integrated storage framework that provides large-scale scalable storage by integrating a plenty of personal cloud storage accounts. Integrating the accounts of different cloud storage providers to deliver a unique access interface makes sense and is quite important. In the proposed storage framework, a plenty of personal free accounts are integrated in order to realize large-scale free personal cloud storage. A prototype system called GrandStore is designed and implemented, which is based on the principle of OAuth protocol and open API. Specifically, after authorized by the owner of account, GrandStore could manage and control the account, so there is no need for complex login any more. Users only need apply for many free cloud storage accounts, and then account authentication credentials are stored in back-end database of GrandStore, which realizes easily enlarging personal free storage space.

Compared with other similar systems, GrandStore is different in three ways. First, GrandStore is a scalable and open system, that is to say, you can add dynamically new accounts to GrandStore without disturbing it. Second, if the developers learn SDK provided by a new personal cloud storage providers, this new product can also be added dynamically to GrandStore. Third, GrandStore depends on database to store user’s authentication credential so as to avoid account login, therefore it can store a plenty of accounts to obtain large space. Since GrandStore has such good features, it is a promising system that has great practical value.

The rest of the paper is organized as follows. Section 2 surveys personal cloud storage system and OAuth account authorization protocol. Section 3 introduces the architecture of GrandStore prototype system. Section 4 describes the implementation of GrandStore prototype system and the final section offers concluding remarks.

II. BACKGROUND AND RELATED WORK

In this section, we introduce the background knowledge about free personal cloud storage, as well as the comparison of several free personal cloud storage, and also introduce cloud storage open platform and OAuth protocol.

A. Free Personal Cloud Storage

Cloud-based services have been introduced in recent years, offering people and enterprises computing and storage capacity on remote data-centers and abstracting away the complexity of hardware management. As one kind of cloud storage, free personal cloud storage has attracted our attention since these years. As the development and popularity of cloud computing, Hadoop Distributed File System (HDFS) has been the first choice to build reliable cloud storage. The comparison of several popular personal cloud storage providers is shown in Table 1, including Amazon Simple Storage Service (S3), Google Drive, Microsoft SkyDrive, Apple iCloud, Dropbox, etc. It is summarized as follows:

- Most of them provide API interface and programming languages support, such as Java, C++, Python, Ruby, C#.

- Most of them provide APIs Client Library for developers.

- Most of them support OAuth 2.0 or 1.1 protocols.

- Most of them provide file synchronization service.

- Most of them provide limited free space for individuals, and there is also file size limit for upload or store. To remove this restriction, you may upgrade to a paid account which will allow you to upload larger files.

B. Open Platform

From the survey on current personal cloud storage providers, we found that they follow the same principle of open platform. Personal cloud storage system open platform allows developers to create their applications to use accounts space, without account login. As it can be seen in Fig. 1, the principle of OAuth[1] account authorization in personal cloud storage open platform is described as the following five steps.

- Step 1: The developer creates an *application*, usually through web page to give the name and some other basic information.

- Step 2: Open platform returns the *application_key* and *application_secret* to developer.

- Step 3: Since the application needs the authentication credential of account, the owner of account is guided to input *username* and *password* to apply for *oauth_token*, which is also called *authorization code* or *access ticket*.

- Step 4: Open platform returns the *oauth_token*.

- Step 5: The developer collects and stores *oauth_token* for further process.

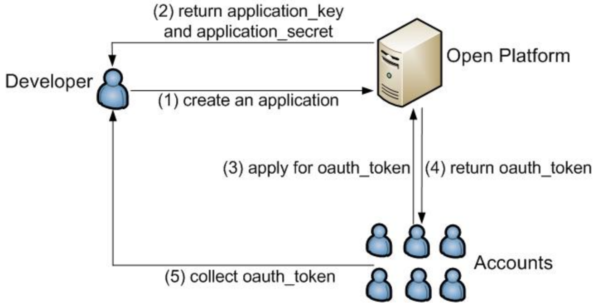


Figure 1. The principle of OAuth-based account authorization cloud storage open platform.

In general, *oauth_token* (or we say *access_ticket*) is composed of two parts, *access_token* and *refresh_token*. Usually, *access_token* has a lifetime, and when it is expired, *refresh_token* is used to generate a new *access_token* and *refresh_token* pair. The expiry period is different, e.g., for some products, it is two weeks; while for some products, it is one month.

C. Related Work

Personal cloud storage services are gaining popularity. From the viewpoint of taxonomic, personal cloud storage belongs to the public cloud filed. Many studies have been reported on personal cloud storage or public cloud storage topic in recent years [1][2][3][4][5][6]. For example, Drago et al. [7] studied the characterization of Dropbox, the leading and widely-used personal cloud storage system, and presented a network traffic measurement and analyzed possible performance bottleneck caused by current system architecture and the storage protocol of Dropbox. In [8], the authors presented the architecture for a secure data repository service designed on top of public clouds to support sharing multi-disciplinary scientific datasets.

In [9], the authors examined the efficacy of leveraging web-based email services to build a personal storage cloud, and then presented EMFS, email-based personal cloud storage, which aggregates back-end storage by establishing a RAID-like system on top of virtual email disks formed by email accounts. In particular, by replicating data across accounts from different service providers, highly available storage services can be constructed based on already reliable, cloud-based email storage, while EMFS cannot match the performance of highly optimized distributed file systems with dedicated servers.

The idea of GrandStore is integrating free personal cloud storage accounts. Similarly, DepSky [10] is a system that provides dependable and secure storage in the cloud through the encryption, encoding and replication of the data on diverse clouds that form a cloud-of-clouds. The authors also deployed the system using four commercial clouds to study the performance.

There are also some approaches and systems proposed recently in hybrid storage or integrated storage area. In [11], the authors proposed a scalable, configurable and reliable hybrid storage system, which is composed of stable volunteered personal cloud storage and P2P-based desktop storage system. BitDew [12] is an open source data management middleware for cloud, grid and desktop grid, developed by INRIA, France. It supports using multi-protocols to transfer files. It can be also used as a management tool to distribute/write files to different nodes using FTP, HTTP, and BT protocols. In [13], the authors presented personal storage grid architecture, which provides end-users with web service interface to allow users consume several cloud data space resources, such as online email account space resource and virtual disk space (e.g. FTP service).

[1] http://en.wikipedia.org/wiki/OAuth

Soares et al. [14] presented the FEW Phone File System, a data management system that combines mobile and cloud storage for providing ubiquitous data access. The system takes advantages of the characteristics of mobile phones for storing a replica of a user's personal data to provide high data availability. Defrance et al. [15] presented the view of home networking as a distributed file system, and proposed a solution to organize the home network according to a gateway-centric architecture, where the content access unification for various devices (UPnP/DLNA devices, personal computers, cloud storage systems, etc) is realized at the file system level.

While MetaCDN [16] works in another way, which proposed harnessing storage clouds for high performance content delivery. Several storage clouds are integrated in MetaCDN, providing a unique access interface. The price and bandwidth of storage clouds are considered, which is used for store decision.

III. System Architecture

The objective of GrandStore is to integrate many accounts to obtain large-scale free space, providing a unique access interface. Users firstly register a lot of accounts, integrate them by GrandStore, and then utilize the unique storage space through GrandStore. Users can manage, integrate, and maintain all their own accounts through GrandStore. The system architecture of GrandStore is shown in Fig .2. As you can see, GrandStore is located in the 'middle layer'. The entities in this figure are explained as follows:

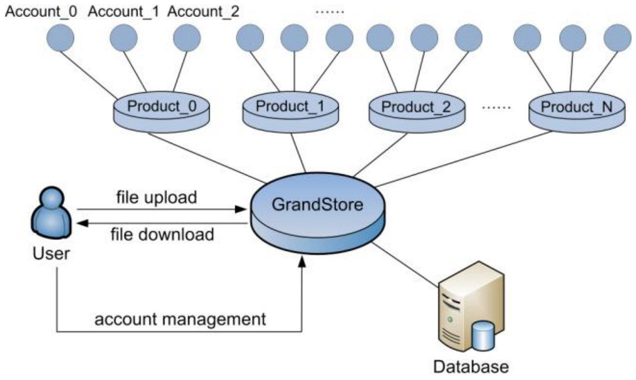


Figure 2. The architecture of GrandStore system.

- Product. It means free personal cloud storage provider, which provides SDK for developers, such as Google Drive, Dropbox and Kuaipan.

- Account. User is allowed to register many accounts for one product. Each account has a limited storage space.

- User. User is responsible for adding dynamically new products and new accounts, and has the right of all kinds of file operations.

- Database. It is used to store account authentication credential (e.g., *access_token*, *refresh_token*) and user's file information (e.g., file name, file size, file path, etc).

GrandStore is based on the principle of OAuth protocol and open API. Specifically, after authorized by the owner of account, GrandStore could manage and control the account, so there is no need for complex login any more. It only needs applying for many free accounts, and then store account authentication credentials to back-end database of GrandStore system, which realizes easily enlarging personal free storage space.

IV. Algorithm and Implementation

GrandStore is released as an open source project, which is developed by Java language, and MySQL 5.0 is selected as the back-end DBMS, and Eclipse is adopted as the development tool. It requires Java SDK provided by personal cloud storage providers. GrandStore now only supports Amazon S3, Google Drive, Dropbox and KingSoft Kuaipan, and it will support more products in the next version.

In this section, we mainly introduce the implementation of core algorithms in GrandStore system. With the aspect of account maintain, we describe account insert algorithm and account authentication credential update algorithm. With the aspect of file operations, we only describe file list, file upload and file download algorithm as three examples. The implementation of other file operations follows the same approach, which is ignored in this paper.

A. *Account Insert Algorithm*

As we mentioned before, we create one application for each product, and we distinguish them through unique application id. For each product, we can also register many accounts, and each account also has a unique id. The account insert algorithm is shown in Algorithm 1. First, GrandStore starts account authorization guide utility. After authorized by the owner of account through inputting username and password, access token and refresh token are generated, and then stored in database. Using this authentication credentials, there is no need for the owner of account to authorize anymore.

Algorithm 1. Account insert algorithm in GrandStore

```
Require: Let $app_i$ be the id of personal cloud storage product
Require: Let account<username, password> be the account to be added
Require: Let $acc_j$ be the id of the account to be added
Require: Let access_token be the returned access token by open platform OAuth server
Require: Let refresh_token be the returned refresh token by open platform OAuth server
Require: Let creation_time be the creation time of authentication credential

Get $app_i$ that account<username, password> belongs to
Start account authorization guide utility for product $app_i$
Input username and password of account
Login and allow $app_i$ to manage the storage space of account
Return authentication credential composed of access_token
  and refresh_token
Get creation_time of this authentication credential
Write {$app_i$, $acc_j$, access_token,refresh_token, creation_time}
  to database
```

B. Account Authentication Credential Update Algorithm

Generally, authentication credential has a lifetime, it will became invalid when exceeds expire period. Therefore, we adopt a multi-thread approach to check all accounts, and find those expired access token. Then, the corresponding refresh token is used to generate a new pair of access token and refresh token, supported by OAuth 2.0 protocol. The detailed algorithm is shown in Algorithm 2.

Algorithm 2. Account authorization credential update algorithm in GrandStore

```
Require: Let $AuthTable\{app_i, acc_j, access\_token, refresh\_token, creation\_time\}$ be the account authentication credential information in database
Require: Let $app_i$ be the id of personal cloud storage product
Require: Let $acc_j$ be the id of account
Require: Let $lifetime_i$ be the lifetime of authentication credential for product $app_i$
Require: Let $current\_time$ be the current time
Require: Let $access\_token_{new}$ be the new access token
Require: Let $refresh\_token_{new}$ be the new refresh token
Require: Let $creation\_time_{new}$ be the creation time of the new authentication credential

for all record $auth \in AuthTable$ do
    Check $auth.app_i$ and get $lifetime_i$ for this product
    if ($current\_time$ - $auth.creation\_time$) > $lifetime_i$ then
        {authentication credential is expired, use refresh token
        to get a new one}
        Check $auth.app_i$ and select corresponding API
        {create an API session for further API calls}
        $API.init(auth.app_i)$
        $API.create(auth.access\_token)$
        $\{access\_token_{new}, refresh\_token_{new}\} \leftarrow$
                        $API.doRefresh(auth.refresh\_token)$
        Get $creation\_time_{new}$ of the new authentication credential
        {update the authentication credential of account $acc_j$}
        Write $\{app_i, acc_j, access\_token_{new}, refresh\_token_{new},$
                $creation\_time_{new}\}$ to database
    end if
end for
```

C. File List Algorithm

Because GrandStore is designed to integrate many accounts, when the user logs into GrandStore, GrandStore should retrieve and then list all files of each account in a unique file access graphical interface. In order to list all the files from all accounts, it just simply executes API calls to get the files of each account. Algorithm 3 describes the file list algorithm.

Algorithm 3. File list algorithm in GrandStore

```
Require: Let $AuthTable\{app_i, acc_j, access\_token, refresh\_token, creation\_time\}$ be the account authentication credential information in database

{list files of each account}
for all record $auth \in AuthTable$ do
    Check $auth.app_i$ and select corresponding API
    {create an API session for further API calls}
    $API.init(auth.app_i)$
    $API.create(auth.access\_token)$
    $API.listAllFiles()$
end for
```

D. File Upload Algorithm

When users upload a file to GrandStore, it firstly lookups a proper account to store this file. In this paper, we propose the ‘maximal unused space’ approach. That is to say, GrandStore lookups the account which has the maximal unused space, and then stores the file to this account. File upload algorithm is demonstrated in Algorithm 4. This approach can achieve storage balance, and avoid the situation that some accounts are too busy than others.

Algorithm 4. File upload algorithm in GrandStore

```
Require: Let $app_i$ be the id of personal cloud storage product
Require: Let $acc_j$ be the id of account
Require: Let $AuthTable\{app_i, acc_j, access\_token, refresh\_token, creation\_time\}$ be the account authentication credential information in database
Require: Let $AccountTable\{acc_j, total\_space, unused\_space\}$ be the account space consumption information in database
Require: Let $max\_space$ be a variable to store maximal unused space
Require: Let $opt\_account$ be the account that has the maximal unused space
Require: Let $F$ be the file to be uploaded to the system
Require: Let $fid$ be the unique id of the file when it is successfully uploaded
Require: Let $FileTable\{fid, acc_j\}$ be the file storage mapping information in database

{lookup the account that has the maximal unused space}
$max\_space \leftarrow 0$
for all record $account \in AccountTable$ do
    if $account.unused\_space > max\_space$ then
        $max\_space \leftarrow account.unused\_space$
        $opt\_account \leftarrow account$
    end if
end for
{lookup the authentication credential of $opt\_account$}
for all record $auth \in AuthTable$ do
    if $auth.acc_j == opt\_account$ then
        {upload to this account directly}
        Check $auth.app_i$ and select corresponding API
        {create an API session for further API calls}
        $API.init(auth.app_i)$
        $API.create(auth.access\_token)$
        $fid \leftarrow API.doUpload(F)$
        {update account space consumption of $opt\_account$}
        $opt\_account.unused\_space \leftarrow$
                $opt\_account.unused\_space - F.size()$
        {insert file storage information}
        Write $\{fid, opt\_account.acc_j\}$ to database
        break
    end if
end for
```

E. File Download Algorithm

File download algorithm is relatively easier than file upload algorithm. When users download a file from

GrandStore, it firstly lookups the account that contains this file, and then download from this account directly through API calls. Algorithm 5 indicates file download algorithm.

Algorithm 5. File Download Algorithm in GrandStore

```
Require: Let $AuthTable\{app_i, acc_j, access\_token, refresh\_token, creation\_time\}$ be the account authentication credential information in database
Require: Let $FileTable\{fid, acc_j\}$ be the file storage mapping information in database
Require: Let $F$ be the file to be downloaded from the system
Require: Let $store\_account$ be the account that stores $F$

{lookup the account that stores $F$}
for all record $file \in FileTable$ do
  if $file.fid == F.getfid()$ then
    $store\_account \leftarrow file.acc_j$
    break
  end if
end for
{lookup the authentication credential of $store\_account$}
for all record $auth \in AuthTable$ do
  if $auth.acc_j == store\_account$ then
    {store in this account, download directly}
    Check $auth.app_i$ and select corresponding API
    {create an API session for further API calls}
    $API.init(auth.app_i)$
    $API.create(auth.access\_token)$
    $API.doDownload(F)$
    break
  end if
end for
```

V. Conclusion

Based on the study of current cloud storage system open platforms and OAuth protocol, this paper proposed a method to integrate a plenty of free accounts to get a unify large-scale free personal cloud storage, and also introduced the design and implementation of a prototype system called GrandStore. The core algorithms of GrandStore are described in detail. It is a promising system that has great practical value. First, it proposed a method to get large storage space without upgrading to a paid account. Second, it allows you to manage all your accounts in a unique access interface.

In spite of this, we plan to improve GrandStore in three ways in future work,

- First, we will improve it, and release a new version for Tablet and Android equipment, to manage your own accounts in a mobile terminal.

- Second, we will design optimized algorithms which consider network distance. Generally, free personal cloud storage providers are geographically dispersed, e.g., the server of Google Drive locates in US, while the server of KingSoft Kuaipan locates in China. When users write or read files, selecting the 'closest' provider or account makes sense and very important.

- Third, we will do some I/O performance evaluation for GrandStore, as well as the advantages of network-aware account selection algorithm.

Acknowledgment

This work is supported by the Natural Science Foundation of Hunan Province under grant no. 2015JJ3071, as well as the Scientific Research Fund of Hunan Provincial Education Department under grant no. 16C0643 and 12C0121.

References

[1] D. Dai, W. Zheng and T. Fan, "Evaluation of personal cloud storage products in China," Industrial Management and Data Systems, Vol 117, Issue 1, 2017, pp. 131-148.

[2] H. Chen, L. Zhang, B. Hu, S. Long and L. Luo, "On Developing and Deploying Large-File Upload Services of Personal Cloud Storage," Proceedings of 2015 IEEE International Conference on Services Computing (SCC 2015), New York City, NY, USA, pp. 371-378.

[3] R. Pitchai, S. Jayashri and J. Raja, "Searchable Encrypted Data File Sharing Method Using Public Cloud Service for Secure Storage in Cloud Computing," Wireless Personal Communications, vol. 90, no. 2, 2016, pp.947-960.

[4] E. Bocchi, I. Drago and M. Mellia, "Personal cloud storage: Usage, performance and impact of terminals," Proceedingd of the 4th IEEE International Conference on Cloud Networking (CloudNet 2015), Niagara Falls, ON, Canada, 2015, pp. 106-111.

[5] K. Ning, Z. Zhou and L. Zhang, "Leverage Personal Cloud Storage Services to Provide Shared Storage for Team Collaboration," Proceedings of IEEE International Conference on Services Computing (SCC 2014), Anchorage, AK, USA, 2014, pp. 613-620.

[6] M. Nebeling, M. Geel, O. Syrotkin, M. C. Norrie, "MUBox: Multi-User Aware Personal Cloud Storage," Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI 2015), Seoul, Republic of Korea, 2015, pp. 1855-1864.

[7] I. Drago, M. Mellia, M. Munafo, A. Sperotto and R. Sadre, A. Pras, "Inside dropbox: understanding personal cloud storage services," Proceedings of the 2012 ACM conference on Internet measurement conference (IMC'12), 2012, pp. 481-494.

[8] A.G. Kumbhare, Y. Simmhan and V. Prasanna, "Designing a secure storage repository for sharing scientific datasets using public clouds," Proceedings of the second international workshop on Data intensive computing in the clouds (DataCloud-SC'11), 2011, pp. 31-40.

[9] J. Srinivasan, W. Wei, X. Ma and T. Yu, "MFS: Email-based Personal Cloud Storage," Proceedings of the 6th IEEE International Conference on Networking, Architecture and Storage (NAS'2011), 2011, pp. 248-257.

[10] A. Bessani, M. Correia, B. Quaresma, F. Andr é and P. Sousa, "DepSky: dependable and secure storage in a cloud-of-clouds," Proceedings of the Sixth European conference on Computer systems (EuroSys 2011), 2011, pp. 31-46.

[11] B. Tang and G. Fedak, "Analysis of data reliability tradeoffs in hybrid distributed storage systems," Proceedings of the 17th IEEE International Workshop on Dependable Parallel, Distributed and Network-Centric Systems (DPDNS 2012), 2012, pp. 1540-1549.

[12] G. Fedak, H. He and F. Cappello, "BitDew: A data management and distribution service with multi-protocol file transfer and metadata abstraction," Journal of Network and Computer Applications, vol. 32, no. 5, 2009, 961-975.

[13] M.-G. Lim, S. Wu, T. Simon, M. Rashid and N. Helian. "Personal Storage Grid Architecture: Consuming Cloud Data Space Resources," International Journal of Grid and High Performance Computing, vol. 2, no. 3, 2010, 17-30.

[14] J. Soares and N. Pregui ça, "Combining Mobile and Cloud Storage for Providing Ubiquitous Data Access," Proceedings of the 17th International Conference on Parallel Processing (Euro-Par 2011), Lecture Notes in Computer Science (LNCS), Volume 6852/2011, Springer-Verlag, 2011, pp. 516-527.

[15] S. Defrance, R. Gendrot, J. Le Roux, G. Straub and T. Tapie, “Home Networking as a Distributed File System View,” Proceedings of the 2nd ACM SIGCOMM workshop on Home networks (HomeNets’11), 2011, pp. 67-72.

[16] J. Broberg, R. Buyya and Z. Tari, “MetaCDN: Harnessing ‘storage clouds’ for high performance content delivery,” Journal of Network and Computer Applications, vol. 32, no. 5, 2009, pp. 1012-1022.

Adaptive Correcting Strokes Extracted From Chinese Characters in Digital Ink of Non-Native Writers Based on Comprehensive Visualization

Hao Bai[1, 2], Xiwen Zhang[1]
[1]College of Information Science, [2]Advanced Chinese Training
Beijing Language and Culture University
Beijing, China
E-mail: {baihao, zxw}@blcu.edu.cn

Abstract—The correcting process for strokes extracted from Chinese characters is the necessary step to extract the errors of writing errors automatically. Visualization of extracted strokes is the prerequisite for manual correction. Therefore, visualization and adaptive correction methods are proposed. To reduce the cognitive burden of correcting, color, brightness, saturation and order number is comprehensively used to visualize extracted strokes. And tag list is applied for correcting different types of extracted strokes, which provides the training set for error extraction classifier in future work. After experimental verification, the method is effective in operational complexity and efficiency.

Keywords-Correction; Visualization; Digital ink; Stroke extraction; Chinese character

I. INTRODUCTION

In the field of computer-assisted Chinese teaching, many scholars have been exploring in recent years [1-3], and developed some can be applied to the actual teaching system of Chinese characters teaching [4-8], but whether from the implementation method or teaching performance, the effect is slightly less. In general, the computer system for the basic process of teaching Chinese characters for the user (especially studying Chinese as the second language) using handwriting input device to enter the Chinese characters into the computer system, the system firstly recognizes the handwriting character to the text character and then find the corresponding template character in the template library. After extracting and matching strokes of the writing and the template character, according to the matching results, writing errors are classified and recognized. Finally, writing feedback is given. In the whole process above, the computer needs to analyze the writing error from extracted strokes. As a result, the corrected training data is necessary for the training of the error classifier. It is impossible for the computer system to completely match all the strokes of all the handwriting characters; so in the production of training data, adaptive correction method is applied to correct different types of extracted strokes. Visualization is the prerequisite to manual correction. Improved visualization could reduce cognitive burden effectively and efficiently. The correction procedure requires simple operation, accurate tag and rich expression including correction information such as many types of writing errors.

There are few researches on the correction method of extracted stroke results at present. In terms of visualization, the color approach is used to visualize the extracted strokes [9]. However, when the number of strokes increasing, the color of strokes is repeatedly rendered that produces more cognitive burdens. In addition, only colors cannot fully express the information of extracted results, such as order information of strokes. The correction method mainly divided into gesture and tagging correction. Commonly digital ink data gesture correction method [10] utilizes mouse and other input devices to correct object for simple and direct operation; but if the amount of data is large, gesture correction requires higher operation skill and precision that results in the reduction of the correction efficiency. On the other hand, the method of tagging correction can be used to correct data along with tagging information. With the better interactive visualization method, it could achieve higher operational efficiency.

In this paper, the approach of correcting the results of extracted strokes is to use the comprehensive visualization method to reduce the cognitive burden. And then we correct the result of extracted strokes and make the necessary data tag information to provide the prerequisite for writing quality feedback in further work. The rest of this paper is organized as follows: Section 2 introduces the visualization for the result of extracted strokes. Section 3 discusses different types of tagging expression for various extracted strokes results. Sect. 4 presents the operation approach using the tagging list to correct different results of extracted strokes. Sect. 5 describes the experiment results and analysis. Sect. 6 draws conclusion from the experiment.

II. VISUALIZATION OF EXTRACTED STROKES

Visualization of extracted strokes is a prerequisite for manual correction. Intuitive and accurate visualization can effectively reduce the cognitive burden of operators and improve their efficiency of corrective operation. This paper adopts colors, graphics and digital texts to visual express the matching result of strokes in digital ink Chinese character. The use of non-repetitive colors render different Chinese characters strokes, especially for the adjacent ones, of which the colors were distinguishing; the circle graphics is used to mark the direction information of strokes, which is that the black circles represent the start point of the stroke and the white ones represent the end point of the stroke; digital labels show the writing order index of strokes in the Chinese character. According to the Chinese character strokes "from left to right, from top to bottom" writing rules, the digital label will be placed in the stroke of the upper left corner near the bounding box of the stroke, in order to avoid overlapping with the writing strokes. Meanwhile the color of digital label is same to the stroke's, which reduces cognitive burden caused by adjacent strokes. The specific steps of the visualization are shown as follows. Fig. 1 illustrates the result of this visualization.

Step 1: The number of strokes is used to divided the half value of hue spectrum to get the weight of stroke color *W*.

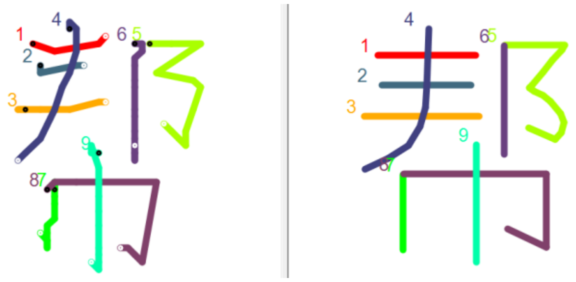
Step 2: Parity of strokes' order is calculated.

Step 3: If the index of stroke is odd, its color is rendered by *HSB(hue, saturation, brightness)* function, of which the value of *hue* is *Index*×*W*, *saturation* is 100% and *brightness* is 100%. *Index* is the index of strokes.

Step 4: If the index of stroke is even, its color is rendered by *HSB(Index*×*W+180, 50%, 50%)*.

Step 5: According to colors of strokes and location of up-left corner of bounding box of strokes, digital indexes of strokes are rendered.

Step 6: Circling the start point of the stroke with black color and the end with white.



(a)the writing charater

(b)the template character

Figure 1. Example of visaulization of extracted strokes

III. TAGGING EXPRESSION OF EXTRACTED STROKES

In the teaching assistant system of Chinese characters, computer system needs to classify the error of writing according to the result of extracted strokes, so as to give the feedback of the Chinese characters handwriting. The training of the writing error classifier requires the correct training data of extracted strokes; but because of the differences in the quality of handwriting and its randomness, it is hardly for computer system to match all the strokes of all character data completely correct. As a result, manual correction is needed to adjust the results of matching strokes as training data. The correction is required as simple and accurate operation but rich of information expression such as types of writing error. Therefore, we use a tag list to reduce the two-dimensional writing data to the one-dimensional array. So that only stroke digital indexes and mark circles remain in the writing panel that solves the problem of information clutter.

A. Defination of Tag List

The approach first defines the data structure of the tag list *Matchlist* includes writing stroke index, template stroke index and mark points coordinate information. The following TABLE 1 illustrates tagging different types of extracted strokes.

TABLE I. ILLUSTRATION OF TAGGING EXPRESSION

Type	Illustration	Examples

one to one	One handwriting stroke correspond to one template stroke	1-1;
		2-3;

concatenated stroke	One handwriting stroke correspond to multiple template strokes	1.1-1;
		1.2-2;
		1.3-3;

broken stroke	multiple handwriting strokes correspond to one template stroke	1.1-1;
		1.2-1;
		1.3-1;

extra stroke	One handwriting stroke correspond to none template stroke	1-0;

redundant stroke	Sub stroke of handwriting correspond to none template stroke	1.1-1;
		1.2-0;

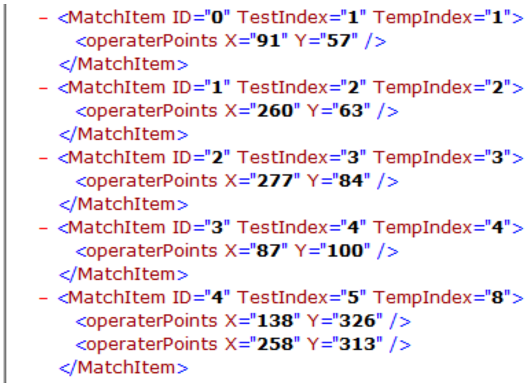
missing stroke	None handwriting stroke correspond to any template stroke	0-1;
		0-4;

B. Recording of Tag List

The method uses the xml file format to store *Matchlist*, which can be used for writing the error classifier after saving the correction information, without directly modifying the original handwriting data and maintaining the integrity of the original data. TABLE 2 illustrates each item in *Matchlist* saved in xml file. Fig. 2 is an example of a saved file fragment.

TABLE II. ILLUSTRATION OF MATCHLIST

Item Name	Illustration	Object Type
MatchItem	Each object is pair of a template stroke and its writing extracted stroke.	MatchItem
ID	Index number of each matchItem	Integer
TestIndex	Index number of strokes in the writing character	Float
TempIndex	Index number of strokes in the template character	Integer
OperaterPoints	Start point and end point in each MatchItem.	Point



```
- <MatchItem ID="0" TestIndex="1" TempIndex="1">
    <operaterPoints X="91" Y="57" />
  </MatchItem>
- <MatchItem ID="1" TestIndex="2" TempIndex="2">
    <operaterPoints X="260" Y="63" />
  </MatchItem>
- <MatchItem ID="2" TestIndex="3" TempIndex="3">
    <operaterPoints X="277" Y="84" />
  </MatchItem>
- <MatchItem ID="3" TestIndex="4" TempIndex="4">
    <operaterPoints X="87" Y="100" />
  </MatchItem>
- <MatchItem ID="4" TestIndex="5" TempIndex="8">
    <operaterPoints X="138" Y="326" />
    <operaterPoints X="258" Y="313" />
  </MatchItem>
```

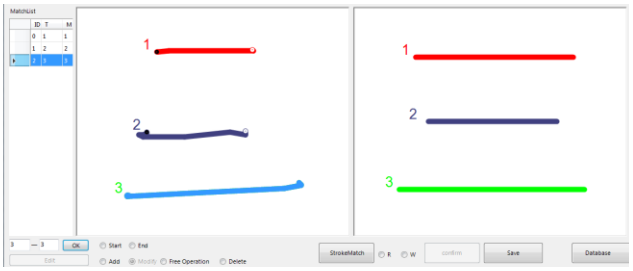
Figure 2. Example of a saved xml fragment

IV. Correction Operation for Extracted Strokes

For the different types of digital ink extracted strokes results, the correction approach proposed in this paper adopts *Matchlist* tagging operation to correct and prepare for the next step in extraction of types of writing errors. The specific operations are show as follows.

A. One to One

- Select the corresponding stroke in the *Matchlist*.
- Click "Edit" to set the stroke start and end point: click "Start" and in the writing panel click for the start point; click "End" and click for the end position in the writing panel on the left side.
- Edit the text box beneath *Matchlist*. The left side is the writing stroke index. The right side is the template stroke index.
- Click "OK" to complete correction, as shown in Fig. 3.



MatchList

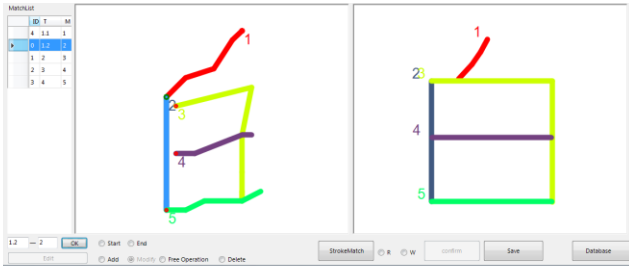
ID	T	M
0	1	1
1	2	2
2	3	3

3 — 3

Figure 3. Example of “One to One” correction

B. Concatenated Stroke

- Select the corresponding stroke in the *Matchlist*.
- Click "Add" to set the stroke start and end point: click "Start" and in the writing panel click for the start point; click "End" and click for the end position in the writing panel on the left side.
- Edit the text box beneath *Matchlist*. The left side is the writing stroke index. The right side is the template stroke index. The input format is "*.*-*" and the integer part of its left side is writing stroke index and the fractional part is written in sequence; the right "*" is its corresponding template stroke index.
- Click "OK" to complete correction, as shown in Fig. 4.



MatchList

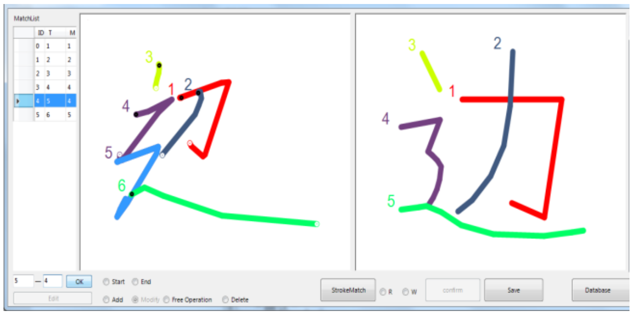
ID	T	M
4	1.1	1
0	1.2	2
1	2	3
2	3	4
3	4	5

1.2 — 2

Figure 4. Example of “Concatenated Stroke” correction

C. Broken Stroke

- Select the corresponding stroke in the *Matchlist*.
- Click "Edit" to set the stroke start and end point: click "Start" and in the writing panel click for the start point; click "End" and click for the end position in the writing panel on the left side.
- Edit the text box beneath *Matchlist*. The left side is the writing stroke index. The right side is the template stroke index.
- Click "OK" to complete correction, as shown in Fig. 5.



MatchList

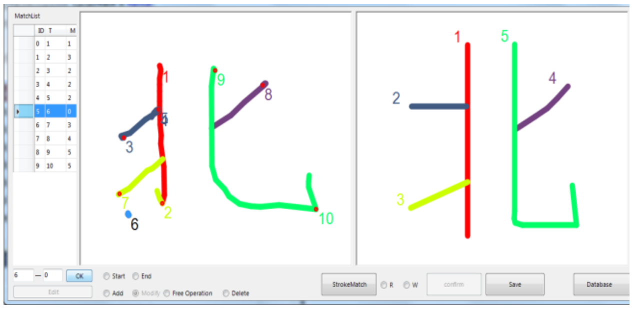
ID	T	M
0	1	1
1	2	2
2	3	3
3	4	4
4	5	4
5	6	5

5 — 4

Figure 5. Example of “Broken Stroke” correction

D. Extra Stroke

- Select the corresponding stroke in the *Matchlist*.
- Click "Edit".
- Edit the text box beneath *Matchlist* as “*-0”. The left side is the writing stroke index. The right side “0” means none of the template stroke corresponding.
- Click "OK" to complete correction, as shown in Fig. 6.



MatchList

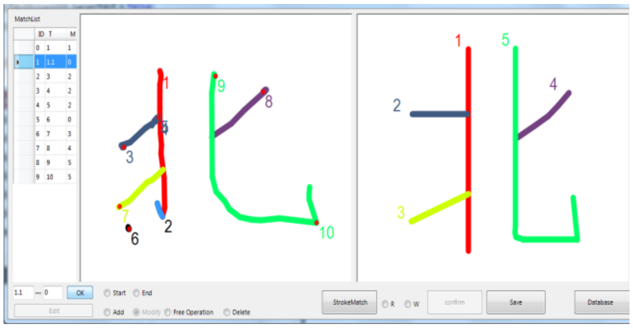
ID	T	M
0	1	1
1	2	3
2	3	2
3	4	2
4	5	2
5	6	0
6	7	3
7	8	4
8	9	5
9	10	5

6 — 0

Figure 6. Example of “Extra Stroke” correction

E. Redundant Stroke

- Select the corresponding stroke in the *Matchlist*.
- Click "Edit".
- Edit the text box beneath *Matchlist* as “*.*-0”. The integer part of its left side is writing stroke index and the fractional part is written in sequence; the right side “0” means none of the template stroke corresponding.
- Click "OK" to complete correction, as shown in Fig. 7.



MatchList

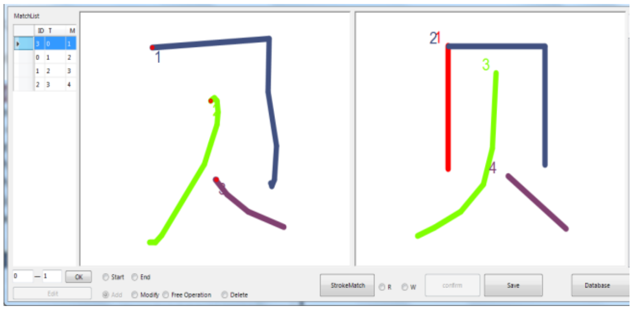
ID	T	M
0	1	1
1	1.1	0
2	3	2
3	4	2
4	5	2
5	6	0
6	7	3
7	8	4
8	9	5
9	10	5

1.1 — 0

Figure 7. Example of “Redundant Stroke” correction

F. Missing Stroke

- Select the corresponding stroke in the *Matchlist*.
- Click "Add".
- Edit the text box beneath *Matchlist* as “0-*”. The left side “0” means none of the writing stroke index corresponding to the template stroke. The right side is the template stroke index.
- Click "OK" to complete correction, as shown in Fig. 8.



MatchList

ID	T	M
3	0	1
0	1	2
1	2	3
2	3	4

0 — 1

Figure 8. Example of “Missing Stroke” correction

V. Experimental Results

The proposed approach is tested in 19815 Chinese characters of 535 kinds including 1094 characters in 6 kinds of errors after strokes extraction, which are realistically handwriting by 127 different foreign Chinese learners [9]. All experiments are run on a PC with Intel Core i7 and 16G RAM. We designed a validation experiment for the proposed correction. The experiment sets the number of keyboard operations and mouse operations for different types of matching results, to verify the complexity of correction. The experiment also records average time of the correction to verify the efficiency. The experimental results are shown in TABLE III.

TABLE III. Results of Experiment

Type	**Number of characters**	**Number of keyboard operations (per character)**	**Number of mouse operations (per character)**	**Time (sec)**
one to one	761	1.1	9.2	28.01
concatenated stroke	133	6.3	17.8	48.83
broken stroke	101	1.2	9.1	22.34
extra stroke	15	1.4	5.1	18.27
redundant stroke	65	4.5	7.3	25.77
missing stroke	19	3.3	6.1	18.38

It can be seen from the experimental results that the proportion of the one-to-one error type is the largest and its correction is within the range of the operation complexity and time consumption. Besides, the number of operation of concatenated stroke is more because strokes are need to splitting extraction, resulting in new strokes, which increases the operation complexity and time consumption.

VI. Conclsion

In the teaching assistant system of Chinese characters, the correction process for matching results of strokes in one character is the necessary step to further extract the types of writing errors intelligently. The visualization of extracted strokes is the prerequisite for manual correction. Therefore, this paper proposes the comprehensive visualization of extracted strokes and adaptive correcting operations. The visualization uses color, brightness, saturation and digital index number to reduce the cognitive burden when correcting. The tagging list is used to express the different types of extracted strokes and prepare for the next type of error type extraction. After experimental verification, the method is effective.

Acknowledgment

This work is supported in part by Science Foundation of Beijing Language and Culture University (supported by “the Fundamental Research Funds for the Central Universities”) (No.17YJ080203), Young and Middle-aged Academic Backbone Support Program of Beijing Language and Culture University and WuTong innovation platform of Beijing Language and Culture University (supported by “the Fundamental Research Funds for the Central Universities”) (No.16PT04). The authors would like to thank Institute of Big Data and Language Education of Beijing Language and Culture University for sharing Skeletal Chinese Character Database.

References

[1] C. B. Zhuang and L. W. Jin, "An Intelligently Verified Algorithm for Correctness and Calligraphy of On-line handwritten Chinese Characters," Signal Processing. vol. 21, Aug. 2005, pp. 276-279.

[2] W. P. Xia and L. W. Jin, "A Method for Layout Evaluation of Online Handwritten Chinese Character Quality Based On Template," Proc. Chinese Conference on Pattern Recognition, 2008, pp. 354-359.

[3] Z. H. Hu, Y. Xu, L. S. Huang, and H. Leung, "A Chinese Handwriting Education System with Automatic Error Detection," Journal of Software, vol. 4, Apr. 2009, pp. 101-107.

[4] J. Li and X. Zhang, "The design and implementation of multimedia intelligent tutoring system for Chinese characters," Proc. IEEE First International Conference on Multi-Media Engineering Education, 1994, pp. 459-463.

[5] H. C. Lam, W. W. Ki, N. Law, A. L. S. Chung, P. Y. Ko, A. H. S. Ho, and S. W. Pun, "Designing CALL for learning Chinese characters," Journal of Computer Assisted Learning, vol. 17, 2001, pp. 115-128.

[6] C. C. Han, C. H. Chou and C. S. Wu, "An interactive grading and learning system for chinese calligraphy," Machine Vision & Applications, vol. 19, 2008, pp. 43-55.

[7] V. Tam and K. W. Yeung, "Learning to write Chinese characters with correct stroke sequences on mobile devices," Proc. International Conference on Education Technology and Computer, 2010, pp. 395-399.

[8] E. Xun, L. Xiaochen, A. N. Weihua, Y. Sun, and I. Ramp, "Stroke Retrieval of Handwritten Chinese Character Images for Handwriting Teaching," Scientiarum Naturalium Universitatis Pekinensis, vol. 51, Mar. 2015, pp. 241-248.

[9] W. An and C. Li, "Automatic matching of character strokes for computer-aided Chinese handwriting education," Proc. International Conference on E-Education, Entertainment and E-Management, 2011, pp. 283-288.

[10] X. W. Zhang, W. H. An and Y. G. Fu, "Adaptive Correction of Errors from Segmented Digital Ink Texts in Chinese Based on Context," Proc. International Conference on Information Technology & Computer Science, 2010, pp. 25-35.

Research on the application of dynamic fuzzy logic in intelligent knowledge base system

Wang Tao
Wuxi Environmental Science and Engineering Research Center,
School of Internet of Things Engineering,
Wuxi City College Of Vocational Technology,
Wuxi, Jiangsu, China
Wang_830@163.com

Abstract—With the Big data under the background of artificial intelligence --AI is increasingly popular, the core role of knowledge base system experts in the increasingly emerge, but whether the automatic driving or artificial recognition need to deal with a lot of expert knowledge data AI, and the expert knowledge not only is fuzzy, and has dynamic. This paper from a new perspective, a comprehensive interpretation of the dynamic fuzzy system theory, and the theory of dynamic fuzzy dynamic fuzzy on the knowledge base of the proposed a new characterization method based on data representation, logical representation. The knowledge base system is also studied and designed.

Keywords-expert knowledge base; data fuzziness; data dynamic; dynamic fuzzy theory

I. INTRODUCTION

In the knowledge base system, the need to deal with a lot of expert knowledge, and the expert knowledge not only is fuzzy, and is dynamic, such as when the financial decisions in the financial aspects of the knowledge base, according to various aspects of the current financial information, combined with the knowledge in the knowledge base construction, to simulate the financial plan select and determine the implementation process of the simulation, the simulation is consistent with the actual degree is fuzzy, and financial solutions in a variety of uncertain financial or economic factors, this ambiguity will reflect the dynamic change, determine again financing plan, the funds required number, capital structure and how / financing channels was reasonable with fuzziness, with the passage of time and all kinds of uncertain factors, the rationality will change, may Become more reasonable or more unreasonable, this ambiguity, dynamic to be able to grasp well, the financial decision-making will be of great benefit, otherwise it will cause greater loss of property. Therefore, in the process of the knowledge base, this paper uses the theory of dynamic fuzzy logic (Dynamic Fuzzy Logic) theory, the knowledge representation, storage, inference and update.

The knowledge base system, based on dynamic fuzzy logic (DFL) of the basic theory, firstly, this paper proposes a new method of dynamic fuzzy knowledge representation method, and the dynamic fuzzy dynamic data into a quantitative structure, the knowledge base is designed.

II. INTRODUCTION TO DYNAMIC FUZZY LOGIC (DFL) THEORY[1] [2] [3]

A. DFL propositional logic

Definition 1 A declarative sentence with dynamic fuzziness (Character of Dynamic Fuzzy) Be called DF proposition (Dynamic Fuzzy proposition)，Capital letters A, B, C..... Express, For a DF proposition (Dynamic Fuzzy proposition), Generally there is no absolute true and false, can only ask it DF true and false (Dynamic Fuzzy or false degree) how?

My daughter, Wang Ruixuan, grew up." Is a DF proposition, "long" reflects the dynamic, "big" reflects the ambiguity.

She is in a better mood." This is also a DF proposition, "turn" embodies the "dynamic", and "good" is fuzzy.

Definition 2 used to measure a DF truth degree by DF number (Dynamic Fuzzy proposition) and($\overleftarrow{a},\overrightarrow{a}$)$\in[0,1]$ to express，The truth of the proposition, Commonly used lower case letters ($\overleftarrow{a},\overrightarrow{a}$),($\overleftarrow{b},\overrightarrow{b}$),($\overleftarrow{c},\overrightarrow{c}$)$\cdots$express。among ($\overleftarrow{a},\overrightarrow{a}$) $=\overleftarrow{a}$ or $\overrightarrow{a}$ max($\overleftarrow{a},\overrightarrow{a}$) $=\overrightarrow{a}$, min($\overleftarrow{a},\overrightarrow{a}$) $=\overleftarrow{a}$ 。

Definition 3 a DF proposition can be considered as a variable on the interval [0,1]，DF propositional variable. For DF variables ($\overleftarrow{x},\overrightarrow{x}$)，($\overleftarrow{y},\overrightarrow{y}$) $\in[0,1]$ The following operations are specified:

①Deny "—"：($\overleftarrow{x},\overrightarrow{x}$)the negation of $\overline{(\overleftarrow{x},\overrightarrow{x})}$， and $\overline{(\overleftarrow{x},\overrightarrow{x})}\triangle=((1-\overleftarrow{x}),(1-\overrightarrow{x}))$

②Disjunctive "∨"：($\overleftarrow{x},\overrightarrow{x}$)∨($\overleftarrow{y},\overrightarrow{y}$) △= max (($\overleftarrow{x},\overrightarrow{x}$), ($\overleftarrow{y},\overrightarrow{y}$))

③Conjunctive "∧"：($\overleftarrow{x},\overrightarrow{x}$)∧($\overleftarrow{y},\overrightarrow{y}$)△= min (($\overleftarrow{x},\overrightarrow{x}$), ($\overleftarrow{y},\overrightarrow{y}$))

④ condition "→"： ($\overleftarrow{x},\overrightarrow{x}$)→($\overleftarrow{y},\overrightarrow{y}$) $\Leftrightarrow\overline{(\overleftarrow{x},\overrightarrow{x})}\vee(\overleftarrow{y},\overrightarrow{y})\triangle=$ max ($\overline{(\overleftarrow{x},\overrightarrow{x})}$, ($\overleftarrow{y},\overrightarrow{y}$))

⑤Double condition “$\leftrightarrow$”：($\overleftarrow{x},\overrightarrow{x}$)$\leftrightarrow$($\overleftarrow{y},\overrightarrow{y}$)△= min(max ($\overline{(\overleftarrow{x},\overrightarrow{x})}$, ($\overleftarrow{y},\overrightarrow{y}$))，max (($\overleftarrow{x},\overrightarrow{x}$) , $\overline{(\overleftarrow{y},\overrightarrow{y})}$))

The definition of propositional formula of 4 DF can be defined as:

①A single DF propositional variable itself is a unified formula;

②if($\overleftarrow{x},\overrightarrow{x}$)PA is unified formula，that $\overline{(\overleftarrow{x},\overrightarrow{x})P}$ Is also a formula;

③if ($\overleftarrow{x},\overrightarrow{x}$)P and ($\overleftarrow{y},\overrightarrow{y}$)Q closed formula，that ($\overleftarrow{x},\overrightarrow{x}$)P ∨ ($\overleftarrow{y},\overrightarrow{y}$)Q，($\overleftarrow{x},\overrightarrow{x}$)P ∧ ($\overleftarrow{y},\overrightarrow{y}$)Q，($\overleftarrow{x},\overrightarrow{x}$)P→($\overleftarrow{y},\overrightarrow{y}$)Q, ($\overleftarrow{x},\overrightarrow{x}$)P$\leftrightarrow$($\overleftarrow{y},\overrightarrow{y}$)Q All are formulas.

B. Predicate calculus of DFL [4][5]

Define recursive definition of 5 DFL predicate formula:

①The atom (first order predicate symbol) is a formula.

②If G, H is a formula, T is the true value of the assigned value of DF, ($\overleftarrow{x},\overrightarrow{x}$) is free variables in DFL，that $\overline{G}$，G ∨ H，G ∧ H，G $\rightarrow$ H，G $\leftrightarrow$ H，($\overleftarrow{x},\overrightarrow{x}$)G，（∀($\overleftarrow{x},\overrightarrow{x}$)G），（∃($\overleftarrow{x},\overrightarrow{x}$)G）is formula.

③All the formulas in DFL are used ①、② for finite times.

Define 6 an interpretation of the formula G in DFL 6 by I and the following rules are composed of U

①Specify a DF element in U for each variable symbol in G;

②Specifying the mapping for each n function symbol in G U T→ D

③Specify the mapping for each n predicate symbol in G DT→ B

Where B is the DF atomic weight, based on these definitions, some properties of the DFL predicate system are listed below

Property 1

$$\overline{(\overleftarrow{T},\overrightarrow{T})\forall(\overleftarrow{x},\overrightarrow{x})G} = (\overleftarrow{1}-\overleftarrow{T},\overrightarrow{1}-\overrightarrow{T})\overline{\forall(\overleftarrow{x},\overrightarrow{x})G} = (\overleftarrow{T},\overrightarrow{T})\exists(\overleftarrow{x},\overrightarrow{x})G$$

Property 2 $$\overline{(\overleftarrow{T},\overrightarrow{T})G} = (1-\overleftarrow{T},1-\overrightarrow{T})G$$

Property 3 $$(\overleftarrow{T},\overrightarrow{T})\forall(\overleftarrow{x},\overrightarrow{x})G = (\overleftarrow{T},\overrightarrow{T})(\forall(\overleftarrow{x},\overrightarrow{x})G)$$

$$((\overleftarrow{T},\overrightarrow{T})\exists(\overleftarrow{x},\overrightarrow{x}))G = (\overleftarrow{T},\overrightarrow{T})(\exists(\overleftarrow{x},\overrightarrow{x})G)$$

III. DYNAMIC FUZZY LOGIC REPRESENTATION OF KNOWLEDGE IN KNOWLEDGE BASE SYSTEM [5] [6]

Constructing knowledge base system is an important and difficult problem. Hundreds of rules and a lot of facts are obtained by visiting experts in the field, and at the same time, the knowledge will be generated during the operation of the model library. The expert knowledge representation into facts and rules is tedious and time-consuming process, the main difficulties are: expert with the way he understands declarative knowledge, these knowledge includes the background, concepts, relations and problems, it is difficult to use a computer program to describe the existence; subjective, uncertain and dynamic problems such as expert knowledge no, the consistency of knowledge including knowledge redundancy, implication, contradictions, omissions and other aspects, this is a problem not to be ignored for the knowledge base system.

A large number of dynamic fuzzy knowledge in knowledge base can be expressed by the method of dynamic fuzzy logic.

A. Dynamic fuzzy logic system[7]

Because the traditional logic system is not easy to deal with dynamic fuzzy problems, here we use dynamic fuzzy logic system, which is composed of a dynamic fuzzy (global) database, dynamic fuzzy logic rules and dynamic fuzzy logic rule interpreter composition.

The dynamic fuzzy database is used to store the initial information provided by the user, the intermediate information obtained in the process of reasoning, and the final conclusion.

The dynamic fuzzy logic rule base is composed of a set of dynamic fuzzy logic rules. A dynamic fuzzy logic rule can be abstractly described as a three tuple::

Prerequisite P, action or conclusion Q, rule of confidence ($\overrightarrow{CF},\overleftarrow{CF}$)) where the preconditions P and conclusion Q can also be dynamic fuzzy

Dynamic fuzzy logic rule interpreter: responsible for part of the rules of the fuzzy and dynamic conditions of the contents of the database according to the rules of credibility ($\overrightarrow{CF},\overleftarrow{CF}$), if the matching success, dynamic fuzzy rule interpreter according to the description of the information content of the action part to modify the dynamic fuzzy database, repeated indefinitely until the issue is resolved.

Dynamic fuzzy logic rules are the traditional DF rules, can be carried out from the following aspects:

①The precondition of DF is to introduce the DF predicate and the DF state quantifier to express the DF relation and the DF state in the rule precondition, and define a DF matching principle.

② Conclusion: the DF action or action or the conclusion of rules has a DF or DF conclusion itself is a predicate or a DF state or action itself is a kind of action to operate with DF DF data in DF database.

③Set rule activation threshold. When the matching degree of the present condition is greater than or equal to the rule, the rule is activated.

④Set rules for reliability ($\overrightarrow{CF},\overleftarrow{CF}$). To determine the credibility of the DFL rules to reflect the degree of credibility, it will somehow affect the credibility of the conclusions or actions.

B. *DFL rule and DF data represen tation method[9] [10] [11]*

First of all, the representation of several DF propositions is given

P=[P’=(A(x) is D), $(\overleftarrow{t},\overrightarrow{t})$]

Here P is a DF proposition, X is the object name, A is the X attribute name, D is a deterministic state expression，P'=(A(x) is D) is P the corresponding deterministic proposition, $(\overleftarrow{t},\overrightarrow{t})$ is to use P'to express the degree DFof P

② P=[A $(\overleftarrow{x},\overrightarrow{x})$ is π $(\overleftarrow{t},\overrightarrow{t})$]

Here π $(\overleftarrow{t},\overrightarrow{t})$ is the membership function of A. $(\overleftarrow{x},\overrightarrow{x})$

③ P=[P’=(A $(\overleftarrow{x},\overrightarrow{x})$ is π $(\overleftarrow{x},\overrightarrow{x})$), $(\overleftarrow{t},\overrightarrow{t})$]

$(\overleftarrow{t},\overrightarrow{t})$ is to use P'to express the degree DFof P

According to the representation of DF proposition, a dynamic fuzzy logic rule

IF(P1,P2,⋯⋯Pm) THEN (Q1,Q2,⋯⋯) WITH $\overrightarrow{CF},\overleftarrow{CF}$
Can be expressed as:

IF
$[(P_1^{'}, f_1,(\overleftarrow{t}_1,\overrightarrow{t}_1))AND(P_2^{'}, f_2,(\overleftarrow{t}_2,\overrightarrow{t}_2))AND\cdots AND(P_m^{'}, f_m,(\overleftarrow{t}_m,\overrightarrow{t}_m))]$
THEN

$[(Q_1^{'},g_1,(\overleftarrow{S}_1,\overrightarrow{S}_1)),(Q_2^{'},g_2,(\overleftarrow{S}_2,\overrightarrow{S}_2)),\cdots]$ WITH ($\overrightarrow{CF},\overleftarrow{CF}$)

Here, P1, P2,...... Pm represents the dynamic fuzzy preconditions of rules， Q1, Q2,...... Dynamic fuzzy conclusions and actions in rules，($\overrightarrow{CF},\overleftarrow{CF}$) express rule strength, $P_1^{'},P_2^{'}$, ⋯ , $P_m^{'}$ is P1,P2,⋯⋯Pm Corresponding deterministic expression. $Q_1^{'},Q_2^{'},\cdots$ is $Q_1,Q_2,\cdots$ Corresponding deterministic expression, $f_1, f_2,\cdots,f_m$ is used $P_1^{'},P_2^{'}$, ⋯ , $P_m^{'}$ expression of P1, P2,...... State probability distribution of Pm. $(\overleftarrow{t}_1,\overrightarrow{t}_1),(\overleftarrow{t}_2,\overrightarrow{t}_2),\cdots,(\overleftarrow{t}_m,\overrightarrow{t}_m)$ is used $(P_1^{'},f_1),(P_2^{'},f_2),\cdots,(P_m^{'},f_m)$ expression of P1,P2,⋯⋯Pm' DF degree.

In this way, the knowledge in the knowledge base system can be represented by the dynamic fuzzy logic.

IV. A NEW METHOD FOR CHARACTERIZING DYNAMIC AMBIGUITY

In theory, the knowledge of dynamic fuzzy degree can be represented by a DF number, but can be found through the analysis, a number of DF actually contains two aspects of fuzzy and dynamic information, is the subject of ambiguity, and put forward the dynamic change trend.

However, in practical applications, the solution of dynamic fuzzy problems is not satisfied with the fuzzy and dynamic changes, but also requires the size of dynamic changes, that is, the degree of dynamic change. In this paper, a new method of dynamic variation is introduced.

Definition: the so-called moment t dynamic change degree D (T), refers to the change in the rate of membership at the time of T, that is, the $f(t)$ derivative of the membership function at the moment

d(t)= $f(t)$

When d (t) >0, the direction of the increase of the value of the degree of membership changes, the greater the value, the faster the speed: when d (T) <0, the direction of the change in the degree of membership value, the smaller the value, the faster the speed.

In practice, it is difficult to obtain the membership function directly, but only a few discrete data:

(1) Difference quotient:

$$d(t)\approx\frac{f(t)-f(t^{'})}{t-t^{'}}$$

Here the moment is usually desirable <t t a moment, that is, the difference between the back of commercial law, of course, can also choose forward difference commercial law, depending on the specific circumstances

(2) Curve fitting method

According to the discrete values of membership degree F, as shown in Table 3.1, the curve fitting, the fitting curve of the membership function y = $f(t)$ And then the fitting curve function in the T derivative is d (t) = $f(t)^{'}$

TABLE I. DISCRETE DATA OF MEMBERSHIP

T_i	t_1	t_2 ⋯⋯ t_m
F_i	f_1	f_2 ⋯⋯ f_m

Curve fitting method is usually applied to the case of a large amount of data.

(3) Expert investigation

For a number of financial management experts, using the questionnaire method to obtain the basic dynamic measurement data, and then get the average dynamic measurement.

This design adopts the backward difference method to get the dynamic fuzzy data dynamic degree.

V. DESIGN OF KNOWLEDGE BASE SYSTEM BASED ON DFL

A. *Functional structure of DFL rule knowledge base subsystem*

The main function of the knowledge base system based on DFL is to provide the knowledge of the whole process. Its main functions are knowledge representation (in this paper, the use of dynamic fuzzy representation), knowledge reasoning (based on dynamic fuzzy logic inference method) and knowledge learning three functions, as shown in Figure 1.

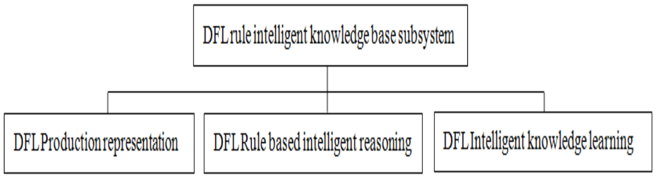

DFL rule intelligent knowledge base subsystem
DFL Production representation
DFL Rule based intelligent reasoning
DFL Intelligent knowledge learning

Figure 1. Function structure of DFL rule knowledge base subsystem

With regard to the method of knowledge learning, FDSS adopts the method of mechanical learning and inductive learning. The mechanical learning method is realized by the maintenance of the knowledge base. The inductive learning method is implemented by running the model in the model base.

VI. SUMMARY

This paper is mainly the theory of fuzzy function and logic design of knowledge base system based on dynamic, which can well solve the representation and reasoning of the dynamic in the objective world and fuzzy problems, so as to deepen and improve the level of intelligent knowledge base system.

ACKNOWLEDGMENT

The author are grateful for financial support from the funding project :

[1] outstanding young teachers and principals outside the training project Department of Education , Jiangsu Province

[2] Education Science Planning Leading Group Office ,Jiangsu Province (grant no.D/2016/03/62)

REFERENCES

[1] Li Fan-zhang, Liu Gui Quan, She Yu-mei etc.. An introduction to Dynamic Fuzzy Logic [M], science and technology publisher in Yunnan, 2005.

[2] Li Fan-zhang The True Value Domain measure of Dynamic Fuzzy Logic;.Computer engineering, 2001,27(3):83-85.

[3] LI Fan-zhang; MEI Yu; QIAN Xu-pei. Research on a Dynamic Fuzzy Data Model. Journal of Chinese Computer Systems, 2002, 23(9):1107-1109.

[4] Qing Pan, Fanzhang Li Theroy of Dynamic Fuzzy Relational De8cision-making and its Applications, 2013,8-18.

[5] Wang Tao , The application of dynamic and obscure logic in intellectual data; Journal of Qinghai Normal University(Natural Science Edition), 2009,2:25-31.

[6] Liu Luo,Guo Li-hong,Software Reliability Growth Model Based on Dynamic FuzzyNeuralNetwork with Parameters Dynamic Adjustment, 2013,02,186-190.

[7] Rui Hui Juang,Dissertation Submitted to Zhejiang University of Technology Degree of Master 2014.09:12-32 .

[8] Wang Tao . Analysis and design of the computer network test system based on UML modeling;Journal of Xin Yu College, 2005.5:26-29.

[9] Meng XF, Ci X (2013) Big data management: concepts, techniquesand challenges. J Comput Res Dev 50:146–169 (in Chinese).

[10] Manyika J, Chui M, Brown B et al (2011) Big data: the nextfrontier for innovation, competition, and productivity. MckinseyGlobal Institute. http://www.mckinsey.com/*/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx. Accessed 03 Sep 2013.

[11] Guo HD (2014) Digital earth: big earth data. Int J Digit Earth7:1–2.

Speaker-dependent Isolated-Word Speech Recognition System Based on Vector Quantization

Yinyin Zhao
Engineering Training Center of Beihua University
Jilin, China
Zhyy8119@126.com

Lei Zhu
Beihua University College of Electrical and Information Engineering
Jilin, China
2295591145@qq.com

***Abstract*—Speaker-dependent speech recognition system requires the system should not only recognize speech, but also recognize the speaker of the segment. In this paper, two indicators are selected—short-time average zero-crossing rate and dual-threshold endpoint to test the signal endpoint through the study of speaker-dependent isolated-word speech characteristics, and MFCC parameters are taken as the characteristic parameters; based on vector quantization, template matching algorithms are designed, and one is adopted to improve LBG algorithm to increase the computing speed; speaker-dependent isolated-word speech recognition system is designed based on vector quantization technique and simulation experiments are conducted in the MATLAB platform under various backgrounds, which proves the system has better recognition effect.**

Keywords-Speech recognition; LBG; MFCC; Vector Quantization;

I. INTRODUCTION

The human speech is the most natural and easiest means of communication in the exchange and transfer of information. Therefore, it becomes a new technique to make machine able to understand human speech and make the communication between human and machine as convenient as the one between human and human, which is explored by people continuously. Speech recognition can be divided into speaker-dependent and speaker-independent speech recognition. Speaker-dependent recognition requires the system not only identify the corresponding speech signal, but also identify the speaker who issues the speech segment[1]. Compared with speaker-independent speech recognition, the speaker-dependent recognition highlights both relevant characteristics of the speech signal, and the personality of speaker. As a result, speaker-dependent speech recognition is widely used in many fields, such as network security, banking systems, stock systems, and security systems.

The application of vector quantization technique in speaker-dependent isolated-word recognition system is mainly studied in this paper, and a simulation study of speech recognition system is conducted, achieving good experimental results.

II. DESIGN OF SPEECH RECOGNITION SYSTEM

There are mainly two parts,that is parameter extraction and pattern matching[2]. The basic components of the speech recognition system are shown in Fig.1. Pretreatment and endpoint detection are to ensure the extracted speech features can reflect the characteristics of the speech signal segment. After the detection of start and end points of the voice segment, the characteristic parameters are extracted, and then the appropriate training algorithm is selected based on the eigenvalues to train them and form template library for pattern matching at the time of speech recognition.

III. EXTRACTION OF SPEECH SIGNAL FEATURES

A. Pretreatment

Pretreatment of speech signals includes three steps: pre-emphasis, framing and windowing[3].

The purpose of pre-emphasis is to emphasize the high-frequency portion of the speech, increasing the high-frequency resolution of the speech, to facilitate spectral analysis or channel parameter analysis, which is usually realized by first-order FIR high-pass filter. The function is:

$$H(z) = 1 - \lambda z^{-1} \tag{1}$$

where λ is set to 0.9375 in the experiment, and Fig.2 shows the comparison of “beihua” before and after the pre-emphasis.

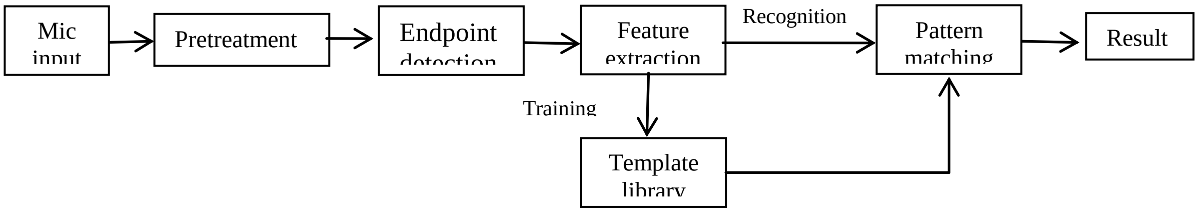


Figure 1. Pre-emphasis before and after comparison

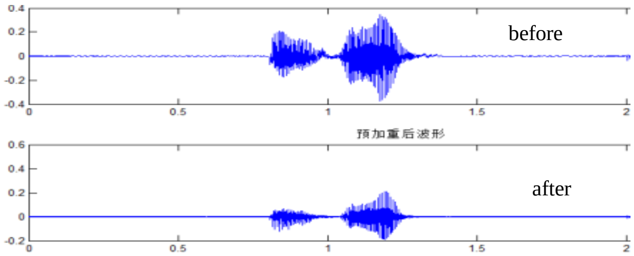


Figure 2. Pre-emphasis before and after comparison

Since the speech signal has the feature of short-time stability, the speech signal can be divided into several frames. According to the sampling frequency of the system, the frame length of system is set as 256, with an overlapped region between two adjacent frames, which makes a smooth transition between frames, ensuring the continuity of signal. This system adopts the frame-shift of 100, and applies Hamming window for the window function.

B. *Endpoint detection*

Characteristics of the speech signal include short-time stability long-time change, and having instant stability, and this time period is typically less than 50ms, so classical stationary signal processing method can be adopted for the processing of speech signal. The traditional endpoint detection is to determine the end through short-time energy and short-time average zero-crossing rate point with short time-domain analysis after the pretreatment of original speech signals, to distinguish pronunciation zone and quiet zone. Short-time energy calculation is carried out based on frame, and short-time energy is defined as[4]:

$$E_n = \sum_{m=n-N+1}^{n} x^2(m) \qquad (2)$$

Zero-crossing rate is an indicator reflecting the signal spectral characteristics. Short-time average zero-crossing rate is the average number of times that waveform crosses zero point within one-frame signal, defined as:

$$Z_n = \sum_{m=-\infty}^{\infty} \left| \mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)] \right| w(n-m) \qquad (3)$$

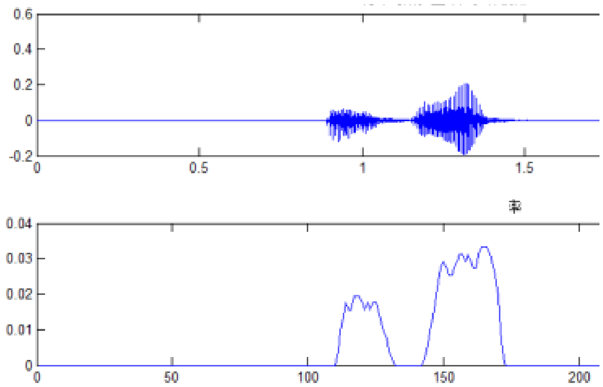


Figure 3. Short-time average zero-crossing rate of “beihua”

Because the amplitude of the speech signal can be reflected in the short-time energy, the frequency is related to the short-time zero-crossing rate, whereas these two indicators can detect sound and silence of signals. In order to detect the start and end points of the speech signal more accurately, dual-threshold endpoint detection algorithm is applied in this paper, which is to combine two indicators (short-time energy and short-time average zero-crossing rate) to detect endpoint detection. Fig.3 is the test result of speech segment “beihua” under the dual-threshold endpoint detection algorithm, in which two lines of the speech signal are the start frame and end frame of the speech segment respectively.

C. *Endpoint Extraction of characteristic parameters*

Extracting characteristic parameters of the speech signal is a key link in the speech recognition process. Characteristics extraction is to analyze and process the speech signal, so as to remove irrelevant redundant information for speech recognition, to obtain the basic characteristic information characterizing human in speech signal. Therefore, the feature information must be able to effectively distinguish different speakers, and keep stable for the changes of the same speaker. There are mainly two feature extraction algorithms used in speech recognition systems currently: Linear Predictive Cepstrum Coefficients (LPCC) and Mel Frequency Cepstrum Coefficients (MFCC)[5-6]. LPCC is an algorithm put forward based on the principles of the human vocalization, while MFCC is proposed based on the human auditory system. Experiments show MFCC has better result than LPCC in speaker-dependent speech recognition, so MFCC is applied to the feature extraction of speech signals in this paper.

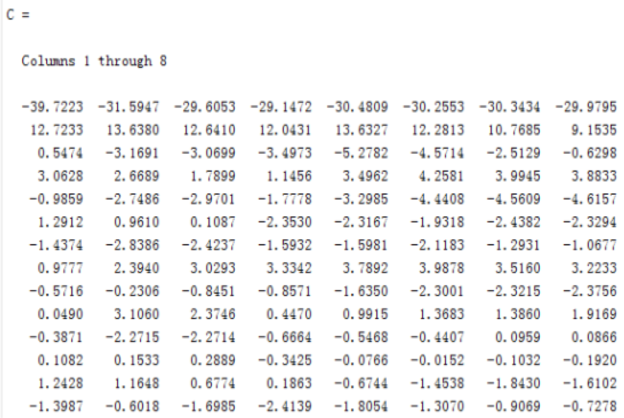
Procedures of MFCC in implementation are as follows:

- Actual frequency is converted into Mel frequency according to the Eq. (4).

$$f_{Mel} = 2595\log_{10}(1+ f / 700) \tag{4}$$

- The output of E (mel) on Mel coordinates passing through this Mel filter group is calculated in Mel frequency;
- The results of the output of E(mel) are transformed into the logarithms by calculation to get logarithmic spectrum S(mel) ;
- The logarithmic spectrum in S(mel) is subject to discrete cosine transform, and then the corresponding MFCC parameters can be obtained. Transformation formula is as follows[7]:

$$C(n) = \sum_{mel=1}^{M} S(mel)\cos\left(\frac{\pi n(mel-0.5)}{M}\right), \quad 0 \le n < M \tag{5}$$



```
C =

  Columns 1 through 8

  -39.7223  -31.5947  -29.6053  -29.1472  -30.4809  -30.2553  -30.3434  -29.9795
   12.7233   13.6380   12.6410   12.0431   13.6327   12.2813   10.7685    9.1535
    0.5474   -3.1691   -3.0699   -3.4973   -5.2782   -4.5714   -2.5129   -0.6298
    3.0628    2.6689    1.7899    1.1456    3.4962    4.2581    3.9945    3.8833
   -0.9859   -2.7486   -2.9701   -1.7778   -3.2985   -4.4408   -4.5609   -4.6157
    1.2912    0.9610    0.1087   -2.3530   -2.3167   -1.9318   -2.4382   -2.3294
   -1.4374   -2.8386   -2.4237   -1.5932   -1.5981   -2.1183   -1.2931   -1.0677
    0.9777    2.3940    3.0293    3.3342    3.7892    3.9878    3.5160    3.2233
   -0.5716   -0.2306   -0.8451   -0.8571   -1.6350   -2.3001   -2.3215   -2.3756
    0.0490    3.1060    2.3746    0.4470    0.9915    1.3683    1.3860    1.9169
   -0.3871   -2.2715   -2.2714   -0.6664   -0.5468   -0.4407    0.0959    0.0866
    0.1082    0.1533    0.2889   -0.3425   -0.0766   -0.0152   -0.1032   -0.1920
    1.2428    1.1648    0.6774    0.1863   -0.6744   -1.4538   -1.8430   -1.6102
   -1.3987   -0.6018   -1.6985   -2.4139   -1.8054   -1.3070   -0.9069   -0.7278
```

Figure 4. Partial MFCC characteristic parameters of the "beihua"

Where M is the order of the Mel filter. M = 24 in this system, and Fig. 4 is a 24-order Mel filter group.

IV. MATCHING AND RECOGNITION OF TEMPLATE

Template matching principle is generally applied for a speaker-dependent small-vocabulary speech recognition system. Firstly, template library is created by the trained speech data, and then feature vectors obtained from the input speech are compared with the templates in the template library, to get the recognition result. Speech recognition algorithms mainly include Hidden Markov Model (HMM), Dynamic Time Warping (DTW), Vector Quantization (VQ)[8], Artificial Neural Networks (ANN) and so on, in which Vector Quantization is a very efficient technique in data compression and coding, which can significantly reduce the computational complexity without reducing the accuracy of identification, simple and practical. Considering the characteristics of the system, the vector quantization method is adopted in the paper to establish the template, which is more suitable for small-vocabulary, isolated-word speech recognition.

A. Vector Quantization

Vector quantization is based on Shannon rate-distortion theory. The theory is that: for a given distorted D, the rate-distortion function R(D) can be calculated.

Where in the average distortion :

$$Q(Y) = \sum_{X} P(X)P(Y/X) \tag{6}$$

Satisfies the condition:

$$\sum_{X}\sum_{Y} P(X)P(Y\mid X)d(X;Y) \le D \tag{7}$$

The inverse function R(D) is the distortion-rate function D(R), which indicates the smallest distortion that the system can achieve under the condition that the given rate is no more than R. Vector quantization is used to increase vector dimension k, and coding performance can be arbitrarily close to the rate-distortion function.

If the input vector is X, of which the dimension is K, then $X = [x_1, x_2, \cdots, x_k]$. The system has two identical codebooks; each codebook contains M codewords Y_i, $i = 1 \sim M$, and each codeword is a K-dimensional vector. VQ encoder principle is to select a corresponding vector Y_i from the encoder codebook according to the input vector X, where the output v is equal to the subscript of the vector, namely, $v = r(X)$. VQ decoder is to select a codeword with the corresponding subscript as output Y according to v by look-up, namely $Y = \beta(v)$.

B. LBG algorithm

LBG algorithm is an efficient and intuitive design algorithm for vector quantization codebook[9]. After the MFCC parameters are extracted, characteristic parameters are trained be the application of basic LBG algorithm to get the corresponding codebook[10]. The basic LBG algorithm is as follows:

All the necessary reference vectors X for VQ codebook training are given, and the set of X is represented by S; quantization levels, distortion control threshold β, maximum number of iterations of the algorithm L and initial codebook Y are established, and the total distortion $D^{(0)} = \infty$; the number of iterations is initialized to 1; the final training codebook is obtained $Y_1^{(m)}, Y_2^{(m)}, \cdots, Y_N^{(m)}$, and the total distortion measure is output as $D^{(m)}$.

The maximum number of iterations distortion L and control threshold β in the algorithm are established in order to avoid infinite loop of iterative algorithm. The value of β is much less than one, and when $\beta^{(m)} \le \beta$, it indicates that the reduction in further iteration calculation distortion is limited, so the calculation can be stopped. L is the parameter to limit the number of iterations in order to prevent the excessive number of iterations when β is set lower.

However, in the basic LBG algorithm, due to the arbitrariness of the initial codebook selection, the problem of empty cell cavity may appear. In order to solve this problem, LBG algorithm of empty cell splitting is adopted in this paper. The main steps of empty cell splitting method is as follows:

- Removing the centroid Y_X from the empty cell;
- Splitting the maximum cell SM, multiplying the centroid YM of SM by the disturbance coefficient $1 \pm \partial$ respectively, to get two codewords, YM1 and YM2, which are taken as references for Voronoi tessellation of two small cells SM1 and SM2。

The advantage of this method is the using of two smallest cells to replace the original large cell, to reduce the quantization distortion, thus improving quantization performance.

C. *Matching of codebook*

As described above, we use cell splitting LBG algorithm to train each of the speech signals to be recognized, and then store the resulted training codebooks separately, to get the corresponding template library. The signals to be identified are subject to a series of treatments to get a codebook when recognition is required, and the codebook is matched with each codebook in this template library, to calculate the distortion measure respectively. If there is a small degree of distortion within a predetermined threshold value L’ (i.e. smaller than the predetermined threshold value L’), it is considered that the speech segment to be recognized matches the template, and the recognition result is obtained to be output in the recognition result site. If all the distortions are greater than L’, there are no matching results in the matching template for the segment of the input speech, and then “No matches” is output in recognition result site. Wherein, considering the complexity of the distortion measure calculation and the implementation simplicity of subsequent hardware, the classical absolute-value average error Euclidean distance measure is applied to calculate the corresponding distortion measure. The absolute-value average error Euclidean distance measure is as follows.

Let x be k-dimensional feature vector of unknown model, y be k-dimensional code vector in the codebook, and x_i and y_i be the components of x and y of the same dimension respectively, then absolute-value average error Euclidean distance measure is defined as:

$$d_1(x,y) = \frac{1}{k}\sum_{i=1}^{k}\left|x_i - y_i\right| \tag{8}$$

D. *Simulation experiment*

According to the theory above, systematic experiment is conducted on the MATLAB platform in the paper. There are totally 6 recording persons in the experiment (three men and three women), who are numbered as Speaker 1 to Speaker 6, and among whom the 4th and the 5th speakers are close in speech feature, with audio sampling frequency 11.025KHz; the speech segments to be measured belong to isolated vocabulary. In the testing process, each speaker pronounces the same word twice, in which one is used for training, and the other for recognition. Fig.5 and Fig.6 shows partial recognition results. Test results shows the statistical result of the recognition accuracy.

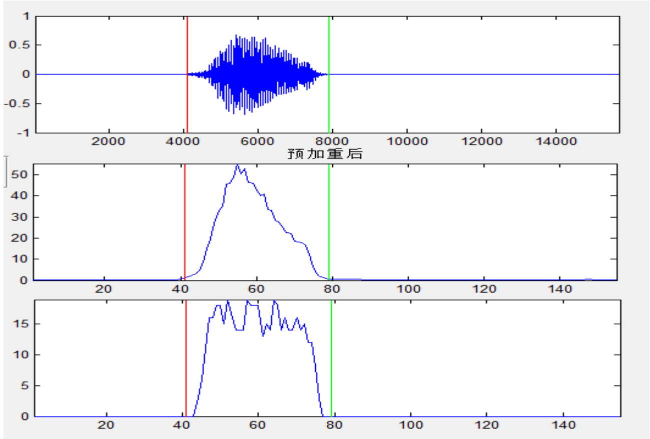


Figure 5. The recognition result of voice "1"

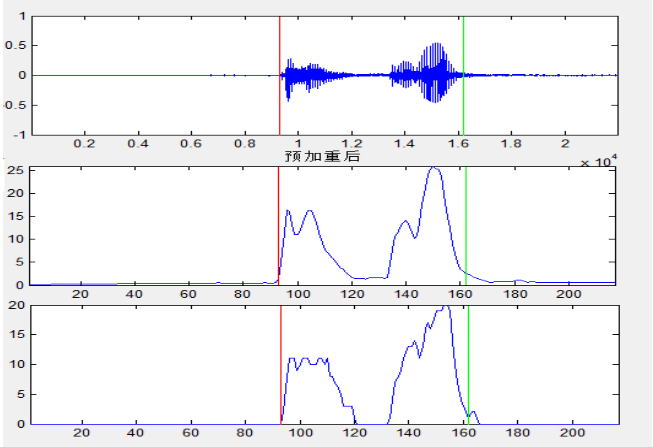


Figure 6. The recognition result of voice "beihua"

From the experimental results, it can be obtained that the system can realize the speaker-dependent isolated-word speech recognition, and the recognition rate can reach 95% in a relatively quiet environment. However, when the speech features of both speakers are close, or environmental noise is great, the accuracy of the recognition will be impacted, and the influence of noise on accuracy is more prominent.

V. CONCLUSION

Vector quantization method is applied in this paper to train the speech to be recognized, and establish the appropriate recognition template library, greatly reducing the amount of computation and data storage, which achieve ideal recognition performance in speaker-dependent isolated-word speech recognition experiment. The treatment of environmental noise is not favorable in this paper, so the next step will be on how to eliminate the influence of ambient noise to improve signal to noise ratio, thus improving recognition accuracy.

ACKNOWLEDGMENT

The authors acknowledge the Science and technology projects in Jilin Province Department of Education (Grant: JJH20170031KJ), the Science and technology projects in Jilin Province Department of Education (Grant: JJH20170035KJ).

REFERENCES

[1] Y. Wang , F. Tang ,J Zheng . Robust Text-independent Speaker Identification in a Time-varying Noisy Environment[J].Journal of Software,2012, 7(9):1975-1980.

[2] Satyanand Singh, E.G Rajan. MFCC VQ based Speaker Recognition and Its Accuracy Affecting Factors[J]. International Journal of Computer Applications. 2011(6) :1-5.

[3] Zhang Y, Long H, Shen S, et al. A novel codebook design with the LBG algorithm in precoding systems under spatial correlated channel[C]. Communications Circuits and Systems (ICCCAS), 2010 International Conference on . 2010:770-775.

[4] Ashkan Parsi, Ali Ghanbari Sorkhi, Morteza Zahedi. Improving the unsupervised LBG clustering algorithm performance in image segmentation using principal component analysis[J].Signal, Image and Video Processing, 2016, 10 (2):301-309.

[5] Chen Chen Huang, Wei Gong, Wen Long Fu, Dong Yu Feng.A Research of Speaker Recognition Based on VQ and MFCC[J]. Applied Mechanics and Materials, 2014, 3468 (644):4325-4329.

[6] M Sahidullah, G Saha. A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition[J]. IEEE Sinal Processing Letters, 2013, 20(2):149-152.

[7] Woo Yong, ChoiHwa Jeon, SongHoon Chung.I-vector Based Utterance Verification for Large-Vocabulary Speech Recognition System[C]. 2016 First IEEE International Conference on Computer Communication and the Internet,2016:334-337.

[8] Lin CY, Prangjarote P, Yeh CH, et al. Reversible joint fingerprinting and decryption based on side match vector quantization[J]. Signal Processing, 2014, 98(1):52-61.

[9] Jian BS.,Robust Multiple Antennas Cooperative Spectrum Sharing Design With Random Vector Qquantization[J]. 2014, 62(4): 486-492.

[10] A Chaudhari,A Rahulkar, SB Dhonde. Combining dynamic features with MFCC for text-independent speaker identification[J]. International Conference on Information Processing,2016:160-164.

Research on Fault Diagnosis Technology of CNC Machine Tool Based on Machining Surface Roughness

Zhou Guang-wen
JILIN ENJINEERINFG NORMAL UNIVERSITY
School of Mechanical Engineering
Changchun, China
zgw_zyl@sohu.com

Tian Mei
JILIN ENJINEERINFG NORMAL UNIVERSITY
School of Mechanical Engineering
Changchun, China
32593829@qq.com

Mao Chun-yu
JILIN ENJINEERINFG NORMAL UNIVERSITY
School of Mechanical Engineering
Changchun, China
290414003@qq.com

Sun Yan-hong
JILIN ENJINEERINFG NORMAL UNIVERSITY
School of Mechanical Engineering
Changchun, China
343175460 @qq.com

*Abstract***—This paper studied the relationship between the spindle fault and the roughness characteristics， by surface roughness of machining. Spindle common fault is divided into the spindle system is not balanced, the spindle system is not right, the spindle system has a transverse crack and the spindle system rolling bearing failure. The characteristic amount of the machining surface is extracted by CCD laser speckle surface roughness measurement technique. Machine fault information and rough surface relationship were established through the adaptive network-based fuzzy inference system（ANFIS）, to achieve the machine tool spindle fault diagnosis. The results indicate that the roughness characteristic can accurately diagnose the machine tool spindle fault and can be an effective method to study the spindle fault of the machine tool.**

Keywords- the spindle fault; roughness characteristics; CCD; ANFIS; machining

I. INTRODUCTION

Spindle as a key power components, widely used in various types of high-speed CNC machine tools, machinery manufacturing plays an extremely important role in the field. With the spindle speed is getting higher and higher, the performance and status of the detection and control has become increasingly important. If the occurrence of degradation, often produce the wrong measurement signal, resulting in improper operation and other serious consequences, and thus CNC machine tool spindle fault diagnosis method has become a more important issue[1].

II. RESEARCH STATUS OF FAULT DIAGNOSIS TECHNOLOGY

Domestic and foreign fault diagnosis technology research shows that the use of fault diagnosis technology, the accident rate and equipment maintenance costs are greatly reduced and improve the productivity. The research results of modern fault diagnosis methods have made great progress in this field. The research results mainly focus on lifting wavelet transform, holographic theory, main fold learning and commonly used diagnostic techniques.

The use of lifting wavelet is very wide, mainly used in signal noise reduction and fault feature extraction. Ana M. Gavrovska and so on for different test signals to select different thresholds, for a purpose of noise reduction and feature extraction, each wavelet filter can be decomposed into the lifting format. ZHANG Yong-xue focuses on the improvement of real signal noise reduction by lifting wavelet transform. Song Guoming uses the lifting wavelet transform to optimize the fault feature, accurately reflect the fault signal characteristic information, and obtain the better fault feature vector through the classification function. Duan Chendong et al. Extract the transient shock fault feature by adding the sliding window based on the lifting wavelet transform. JIANG Quan-sheng et al. Proposed a fault pattern recognition method, using LE algorithm to extract the inherent low-dimensional manifold of the original fault signal, improve fault classification and recognition ability[2,3].

At the same time, other theories are also applied to fault diagnosis, and achieved fruitful research results. Such as particle swarm optimization algorithm, genetic algorithm, support vector machine (empirical mode decomposition, neural network, etc.) have also been very good development and achieved some success, but the application of the machining rough shape to predict the CNC machine tool spindle The fault information is also no one to study.

III. MEASUREMENT OF SURFACE ROUGHNESS OF PARTS

Surface roughness is closely related to the performance of the part. Although the theoretical formula can be obtained through the theoretical formula, but in the actual process, due to cooling and lubrication, tool status, chip and other uncertainties, the theoretical roughness value and the actual value will be different, The consistency of the entire surface of the machined part or all parts roughness. At present, domestic and foreign scholars extensively use a variety of sensors to collect the cutting force, vibration, displacement and current and other signal data during processing, through the establishment of process signal characteristics and surface quality of the approximate relationship model to achieve the measurement of roughness[4].

A. Principle of Three - Dimensional Roughness Measuring Instrument

The digital image processing based on the speckle image is based on the hardware speckle acquisition system. As shown in Fig. 1, the system is a block diagram of the CCD speckle acquisition system. The system is composed of a laser, a CCD camera, a computer and a lens. The computer built-in camera Of the image transmission acquisition card. The laser system is simple, the laser is irradiated to the surface of the standard flat grinding specimen at a fixed angle. The CCD camera collects the speckle image of the standard sample surface in the corresponding fixed position, the computer saves the image transmitted by the camera through the image acquisition card, Spot image for digital image processing.[5]

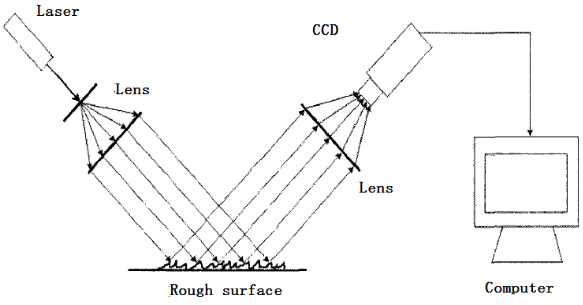


Figure 1. 3D roughness meter acquisition system block diagram

B. Establishment of 3D Roughness Base Plane

The basic principle of three-dimensional evaluation of surface topography is to use a polynomial function to represent the surface to be measured. The function is established on the basis of the least squares principle, and the polynomial function is derived from the derivative, and then the polynomial function is obtained. Coefficient to establish the least squares mid-plane of the contour surface.

The evaluation of the three-dimensional surface roughness parameter is based on the least squares midpoint of the contour, which refers to the surface in which the sum of the squares of the contours of the sampling points in the contour surface is the minimum.

For example, the given surface z (x, y) = f (x, y) can be expressed as polynomial (1):

$$\varepsilon = \sum_{i=1}^{m} \sum_{j=1}^{n} (z(x_i, y_j) - f(x_i, y_j))^2 \quad (1)$$

(m * n is the number of measuring points, i = 0,1, ..., m; j = 0,1, ..., n). Let ε^2 = min, obtain the value of the parameter in the equation, this method of determining the coefficient is called the least squares method. The principle of equivalence is expressed as follows: the most reliable value of the measurement is the square of the variance (the square of the standard deviation), or the square of the uncertainty, or the sum of the squares of the residuals is the minimum. This is a mathematical optimization technique that uses the square of the minimized error to find the best function of a series of data matches.

C. Roughness Feature Extraction

For the same material, the same processing method to generate the surface, the fault information is mainly reflected in the micro-surface texture of the parts and convex and concave, so the accuracy of the parts of the roughness measurement requirements are relatively high, you can extract the surface texture direction Std , The maximum depth of the contour, and the maximum peak height, Sp, for fault assessment. At the same time, the surface root mean square deviation Sq, the surface ten point height Sz, the bottom depth Sv in the parameter selection also need to be considered. The parameters are calculated as follows:

1) the root mean square deviation of the surface Sq: Sq is a frequently used parameter, which indicates the standard deviation of the sample in the statistical field. However, Sq cannot be reflected by the interval distribution and the frequency of the micro-surface deep valleys and the peak The

2) surface ten point height Sz in a sampling area, the surface ten point height represents the depth of the five deepest pits and the arithmetic mean of the height of the five highest vertices. For the three-dimensional evaluation, the definition of different vertices has a great influence on the Sz measurements, but a large number of experiments show that the vertices defined in the relevant region have relatively stable results. However, different sampling area size and location and sampling accuracy of the Sz has a very significant impact, so use it to characterize the stability and effectiveness cannot be effectively guaranteed.

3) Maximum peak height of the contour S p: In the evaluation area D, the maximum value of the z-coordinate on the roughness surface, that is, the maximum height of the contour surface relative to the reference plane. It is also an important parameter that reflects the characteristics of surface roughness.

4) the texture of the surface direction S td in the conventional processing technology processing surface, will produce a certain processing texture, the parameters that the surface texture direction, the metric is relative to the y axis (set the x-axis for the measurement direction) The texture direction of the surface is determined by the surface spectral moment and the cross autocorrelation function. Std is only suitable for the main texture of the surface. In general, the surface of the general accuracy level, there are more obvious texture direction, while the ultra-precision surface, there is no obvious texture direction.

5) Contour maximum peak height S p: In the evaluation area D, the maximum value of the z-coordinate on the roughness surface, that is, the maximum height of the contour surface relative to the reference plane. It is also an

important parameter that reflects the characteristics of surface roughness.

6) the maximum depth of the outline of the depth of Sm in the assessment area D, rough surface z coordinate minimum value, take the absolute value of the contour after the maximum valley depth, that is, to measure the contour surface below the reference surface of the maximum distance. It is another important parameter for the roughness surface feature evaluation, which is echoed with the maximum peak, and has the same important significance[6].

D. ANFIS Model

1)ANFIS structure

For ease of illustration, it is assumed that the ANFIS system with Takagi - Sugeno fuzzy rules has two inputs x and y, one output z, and contains the following two rules.

if x is A1 and y is B1 then f1 = a1x + b1y + c1 (2)

if x is A2 and y is B2 then f2 = a2x + b2y + c2 (3)

The corresponding ANFIS structure is shown in Figure 2The connection between nodes only indicates the flow of the signal, and there is no weight associated with it. The square node represents the node with adjustable parameters, and the round node indicates that there is no tunable node. Among them, only the first layer and the fourth layer has adjustable parameters[7].

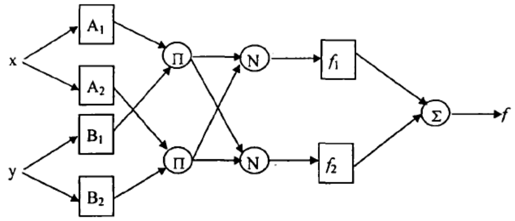


Figure 2. ANFIS structure diagram

The function of each layer follows the first 1-5 layers[4]:

Tier 1 fuzzy input variables, x and y determine the membership of fuzzy sets and the output is

$$o_i^1 = \mu_{A1}(x) \tag{4}$$

According to the chosen form of membership function, you can get the appropriate set of parameters, called the antecedent parameters. All {σi, di} parameter set consisting of the front piece, the learning process in the sample, e.g. the value, select a common Gaussian membership function, the formula (4) can be adaptively adjusted in ANFIS each parameter.

$$\mu_{A1}(x) = \exp\left[-\frac{\left\|x-d_i\right\|^2}{\sigma_i^2}\right] \tag{5}$$

Tier 2 output is the product of the input signal, and its meaning is a sample of the rules activation intensity.

$$o_i^2 = \varpi_i = \mu_{Ai}(\sigma)\mu_{Bi}(y) \tag{6}$$

The fourth layer, go fuzzy computing nodes, each node to calculate the output of the corresponding rules.

$$o_i^3 = \varpi_i = \frac{\omega_i}{\omega_1 + \omega_2} \tag{7}$$

Layer 4 to blur compute nodes, each node to calculate the corresponding.

$$o_i^4 = \varpi_i f_i = \varpi_i(a_i x + b_i y + c_i) \tag{8}$$

Output rulesLayer 5 to calculate the output of all the rules and.

$$o_i^5 = f = \sum_i \varpi_i f_i = \frac{\sum_i \varpi_i f_i}{\sum_i \varpi_i} \tag{9}$$

Structure visible from the ANFIS, ANFIS and fuzzy inference system equivalent. Using a neural network, fuzzy reasoning has the ability to self-learning.

IV. ESTABLISHMENT AND EXPERIMENT OF FAULT DIAGNOSIS MODEL FOR CNC MILLING MACHINE

A. Spindle System Common Faults

1)the spindle system is not balanced. Due to the design, manufacture, installation, processing, spindle long-term operation, electromagnetic interference and other reasons will cause the spindle system quality eccentricity, when the spindle system is running, this eccentricity will cause the spindle system vibration, the spindle system is not balanced. The vibration frequency and the slew frequency are the same. The curve in the time domain is similar to that of the sine wave. The axis of the axis is represented by a circle or an ellipse. If it is a spindle caused by a coupling, the vibration frequency and the rotation frequency are the same. The roughness of the surface can be extracted. The surface texture direction Std has the characteristics of circle or ellipse, the maximum depth of the contour and the maximum peak height Sp of the contour is more than 5% of the standard value.

2)the spindle system is not correct. Spindle system misalignment may be due to manufacturing, installation, support deformation and other reasons, the main form of performance for the coupling is not correct. Couplings are not divided into parallel, angle and both misaligned. The main features are: parallel to the situation will often appear radial vibration, vibration frequency of the frequency of the second frequency of the frequency, but also the existence of multi-frequency vibration; maximum vibration mainly in the coupling on both sides of the bearing, vibration The value increases with the increase of the load; the spectrum shows the radial double frequency and quadruple frequency vibration is obvious. The roughness characteristic can extract the surface texture direction Std radial direction, the surface ten point height Sz is greater than the standard value of 10% or more.

3)the spindle system transverse crack. When the spindle is defective, the lateral crack will be generated on the spindle. When the spindle system is running, the operating state of the spindle system is often affected by the lateral crack. The influence of the opening and closing of the crack on the performance of the spindle system is very serious. When the crack is closed, the vibration characteristic of the main shaft is the same as that of the axis, and the roughness characteristic can extract the surface. When the crack is closed, the vibration characteristic of the spindle is the same as that of the axis. Texture direction Std radial, surface ten point height Sz greater than the standard value of more than 5%.

4)rolling bearing failure. Rolling bearings are generally composed of four basic elements: the rolling body itself (the ball or roller, inner ring, outer ring and cage. When the bearing rotates, these elements interact mechanically; their inherent defects cause the bearing force and The rotation of the rotation axis causes the spindle error to move, and each bearing part has its own shape error, and such shape error will produce the error movement in the spindle. Compared with the basic frequency is the diameter ratio of the bearing element and the rotating element The roughness characteristic can be extracted from the surface texture direction Std is uncertain, the surface ten point height Sz is greater than the standard value of 2-5%.Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you[8-9].

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

B. Fault Mode and Fault Analysis

The fault features E = {E1, E2, E3, E4} are: E1 surface texture direction Std has a circle or ellipse characteristics, the maximum contour depth Sm and contour maximum peak height Sp is greater than the standard value of more than 5%. E2 surface texture direction Std radial, surface ten point height Sz greater than the standard value of 10% or more. E3 surface texture direction Std radial, surface ten point height Sz greater than the standard value of more than 5%. E4 surface texture direction Std is uncertain, the surface ten point height Sz is greater than the standard value of 2-5%. The main reason for these failures F = {F1, F2, F3, F4} are: F1 spindle system is not balanced, F2 spindle system is not right, F3 spindle has cracks, F4 rolling bearing failure. Moreover, the relationship between these phenomena and causes is complex, with obvious nonlinearity and coupling.

C. Fuzzification of Fault and Acquisition of Training Sample s

ANFIS is used to diagnose the NC machine tool. First, the fuzzy rules are determined and the data is fuzzified according to the importance of fault features and cause appearance. The fault feature F1 → F4 is described by two fuzzy sets of "normal" and "high". The feature F4 is represented by two sets of "low" and "normal" modules. According to the expert experience, the trigonometric function is used as the membership of the input variable The function of the membership function of each input variable is shown in Fig.3

Combined with the experience accumulated in the maintenance of the machine tool and the experience of the relevant experts, the fuzzy description of the degree of failure (Table 1) is obtained, and the fuzzy regularization library corresponding to the fault feature and reason is obtained. In order to get a more robust diagnostic model, the fuzzy rules should be as comprehensive as possible.

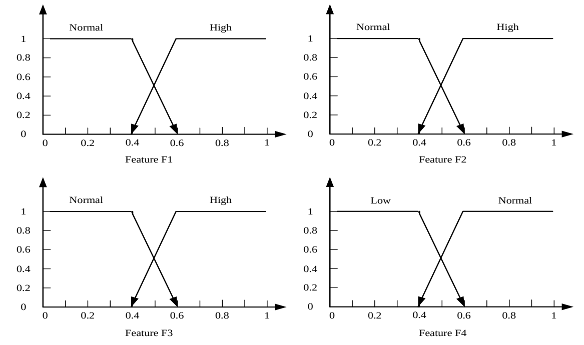


Figure 3. Membership function distribution

TABLE I. TABLE1 FUZZY DESCRIPTION

The Degree of Presence Cause	Membership
Must	0.8～1
Possible	0.6～0.8
Unclear	0.4～0.6
Not too possible	0.2～0.4
NO	0～0.2

TABLE II. SYMPTOM FUZZY RELATIONSHIP WITH THE CORRESPONDING

Fault No.	E1	E2	E3	E4	F1	F2	F3	F4
1	0	0	0	0	0	0	0	0
2	0.8	0	0	0	1	0	0	0
3	0	0.8	0	0	0	1	0	0
…	…	…	…	…	…	…	…	…
29	0	0	0.8	0.8	0	0	0	1
30	0	0.8	0	0.8	0	0	1	1

D. Model Training and Comparison of Results

In the training model, 20 of the 30 groups of samples in Table 2 were randomly selected as training samples and 10 groups as test samples. ANFIS learning was reduced to the adjustment of the front parameter (nonlinear parameter) and the latter parameter (linear parameter) The For the front parameter of equation (5) and the posterior parameters of equation (8), the BP algorithm (back propagation algorithm) and the least squares estimation algorithm are used to adjust the parameters, called the mixing algorithm, according to the relationship between input and output. One of the iterations of the hybrid learning algorithm consists of two steps: Step 1: The parameters of the front part are fixed and the input signal is passed forward along the network until the fourth layer. The least squares estimation algorithm is used to adjust the parameters of the latter[10]. After that, the signal continues along the network Forward to the output layer (ie, layer 5); Step 2, the error signal will be transmitted back along the network, and use the BP algorithm to adjust the front parameters[11]. By using the hybrid learning algorithm, we can get the global optimal point of the posterior parameter for a given preamble parameter, which can not only reduce the dimension of the search space in the gradient descent method, but also can greatly improve the convergence rate of the parameter. It can be seen that the test output only number 5, a group of fault categories cannot be accurately obtained, the remaining five groups of faults can be accurately identified.

TABLE III. DIAGNOSTIC MODEL TEST OUTPUT AND EXPECTED OUTPUT

Fault No.	**Test Output**				**Expected Output**				**Fault Feature**
	C1	***C2***	***C3***	***C4***	***F 1***	***F 2***	***F 3***	***F 4***	
1	0.90	0.10	0.21	0.28	1	0	0	0	E1
2	0.57	0.88	0.16	0.33	0	1	0	0	E1,E2
3	0.22	0.49	0.92	0.16	0	0	1	0	E2,E3
4	0.23	0.29	0.32	0.96	0	0	0	1	E4
5	0.13	0.23	0.85	0.86	0	0	0	1	E3,E4
6	0.37	0.62	0.32	0.86	0	1	0	1	E2,E4

V. CONCLUSION

The fault data of the machine tool spindle are characterized by the extraction of the parameters of the surface roughness. The experimental data prove the effectiveness, but the fault diagnosis accuracy is not accurate enough, and the fault diagnosis model still needs to be further improved.

ACKNOWLEDGMENT

Jilin Province Science and Technology Department funded projects ,China (20160101271JC).

Corresponding author: SUN Yan-hong, JILIN ENJINEERINFG NORMAL UNIVERSITY, School of Mechanical Engineering , Changchun, China.

e-mail:343175460 @qq.com .

REFERENCES

[1] Zhang Jian, based on the image of the workpiece surface roughness detection system research [D], Nanjing University of Aeronautics and Astronautics, 2011.

[2] Xu Xiao Mei, Hu Hong. Development of non-contact surface roughness measurement in last decades [J]. ICMTMA 2009, 1:210-213.

[3] MAO Chun-yu, ZHOU Guang- Wen, XU Yu-kun,Research of Pre-Rotating Machinery Fault Diagnosis Based on Fuzzy Neural Network And Information Fusion, International Symposium on Computer, Consumer and Control, Taiwan, 2014.06.10-12.

[4] Wang Jiahai, Huang Jiangtao, Shen Bin, and so on. Research Status and Prospect of CNC machine tools Intelligent Fault Diagnosis [J]. Machinery manufacturing, 2014 (5): 30 – 32.

[5] FERREIRO S, SIERRA B, IRIGOIEN I, et al. A bayesiannetwork for burr detection in the drilling process [J]. Journalof Intelligent Manufacturing, 2012, 23 (5): 1463-1475.

[6] JOSéVICENTE ABELLáN NEBOT, FERNANDO ROMEROSUBIRON. A review of machining monitoring systems basedon artificial intelligence process models [J]. International Journal of Advanced Manufacturing Technology, 2010, 47 (1/2/3/4): 237-258.

[7] Wang Jianguo, WU Qing, Qin Bo, and so on. Prediction with oxygen [J] PSO support vector machine BOF. Foundry Technology, 2014 (8): 1806-1809.

[8] Wang Jiahai, Huang Jiangtao, Shen Bin, and so on. Research Status and Prospect of CNC machine tools Intelligent Fault Diagnosis [J]. Machinery manufacturing, 2014 (5): 30 – 32.

[9] SUN Yan-jie, AI Chang-sheng.Study on tool wear state monitoring based on fusion of cutting and cutting force parameters [J]. Combined Machine Tool and Automation Processing Technology ,2011 (05).

[10] XIAO Hong-jun.Fuzzy Control and Experiment of Inverted Pendulum Based on Sugeno Model [J]. Journal of Xi'an Engineering University, 2011 (05).

[11] MAO Chun-yu, ZHOU Guang- Wen, TIAN Mei, esearch Early Mechanical Failure of CNC Motorized Spindle Prediction Method Base on D-S Evidence Theory Information Fusion, IMEICI 2016, Shengyang, china, 2016.09.24-26.

Application Research of Virtual 3D Animation Technology in the Design of Human Computer Interface

Zhou Xiaocheng
Animation College
Anhui Xinhua UniversityHefei, Anhui, China
729334333@qq.com

***Abstract*—As everyone knows, the time of virtual reality has come, how to control and correct use of VR has become a significant topic. "In the era of virtual reality better human-computer interaction", "virtual reality interactive context and traditional context where is the difference in space", "touch interactive virtual reality experience what? Through the man-machine interface design, the use of virtual 3D animation technology, implementation of new methods and new ways of product design.**

Keywords-virtual reality;3D animation;human-computer interaction;product design; industrial design

I. THE BASIC PRINCIPLES AND CHARACTERISTICS OF VIRTUAL REALITY TECHNOLOGY

The virtual reality technology involving computer graphics, sensor technology, dynamics, optics, artificial intelligence and social psychology research, is a higher realm of development of multimedia and 3D technology. Virtual reality technology is a kind of immersive interactive environment based on calculable information, is a new man-machine interface. Specifically, it is using modern technology to generate computer technology as the core of realistic visual, hearing, touch the integration of the specific range of virtual environment, users with the necessary equipment in a natural way with the virtual environment in the interaction of such interactions, resulting in the real environment of the same feelings and experience.

The virtual reality technology was developed, people have been interested in it. The virtual reality technology not only has begun in the military, medicine, real estate, design, archaeology, art, entertainment and many other fields has been more and more widely used, but also the society has brought huge economic benefits. Therefore, the industry believes that 1980s is the personal computer era, 90s is the era of multimedia, network, and in twenty-first Century will be the era of virtual reality technology.

Virtual reality technology is an important direction of simulation technology, simulation technology is a collection of computer graphics and human-computer interface technology multimedia technology network technology sensing technology, is a challenging subject of cross technology and research field. It mainly includes the simulation environment, perception, natural skills and sensing equipment. The simulation environment generated by computer, real-time dynamic 3D realistic image.

The virtual reality as a kind of human-computer interaction characteristics of man-machine interface, also can be called "natural human-computer interface. In this environment, users see the full body color scene, hear the sound in the virtual environment, hands or feet can feel the virtual environment back to his forces, thus causing users a feeling that one is personally on the scene. In the same world and feel the way to feel the virtual world created by the computer, and the corresponding real world has the same feeling. The computer world can be beyond the virtual environment we live outside of time and space, can also be a simulation of the real world, can let a person a virtual graphical interface. Feel personally on the scene.

Virtual reality is a kind of machine interface is more ideal person between the computer and the user form. Compared with the traditional computer interface, virtual reality system has three important characteristics: Immersion, Interactivity, Imagination, any virtual reality system can be used in the three "I" to describe the characteristics of the sense of immersion and interactivity. Is the key to decide whether the system is characteristic of a virtual reality system.

Immersion called Pro sound sense. Virtual reality technology is based on human visual, auditory physiological and psychological characteristics, lifelike three-dimensional imagegenerated by the computer, the user through a head mounted display, data glove or data clothing and other interactive device, can put themselves in the virtual environment, to become a member in the virtual environment. The interaction of various objects in the virtual environment and users, just as in the real world. When the user moves the head when the image in the virtual environment in real time to follow the changes, with the mobile object can gesture and movement, can also hear 3D simulation sound. Users in the virtual environment, everything feels very realistic there is a feeling, personally on the scene.

The man-machine interaction in the virtual reality system is almost a natural interaction, users can not only use the computer keyboard and mouse to interact, but also through the special helmet, gloves and other equipment for sensing data interaction. Computer can according to the user's head, hands, eyes, body language and movement, to adjust the image and sound system. Users through their own language, body movement or action such as natural skills, observation or operation of any object in the virtual environment. The virtual reality system with visual, listening, touching, sensing and reaction device, so the user can obtain the kinesthetic, visual and auditory in the virtual environment, tactile, kinesthetic etc. a variety of perception, so as to achieve personally on the scene feeling.

II. THE KEY TECHNOLOGY AND THE RESEARCH OBJECT OF VIRTUAL REALITY

The virtual reality system according to their different functions, can be divided into immersion type virtual reality system, enhance the reality of virtual reality system, desktop virtual reality system and distributed virtual reality system. Four types of virtual reality system with three "I" characteristics, the system mainly includes the basic group into the observer, sensor, composition effect generator and real simulator.

Essentially, virtual reality is an advanced computer user interface, it also provides to the user through such as video, listen, touch and other intuitive and natural real-time interactive means, to maximize the convenience of the user operation, so as to reduce the burden on the user, improve the working efficiency of the whole system. But the real imaginary and the reality of virtual object and high performance computing technology are 3 main aspects of VR technology.

How will the objects and events in the real world to the virtual environment, is a perception problem. Network technology is how to allow multiple users to participate in the same virtual environment. This requires a distributed mapping structure. A kind of virtual reality is the world space to multi-dimensional information space, including the basic model construction, space tracking, sound localization, the key technology of visual tracking and viewpoint induction, generate these technology makes the realistic virtual world, virtual environment detection and operation data for operation of user access is possible.

How is the real reality of virtual object in virtual environment can directly signal adults feel according to (sound, light, electricity). A display (output), but also to ensure that users get the same vision, or similar real environment from the virtual environment the key technology of haptic and auditory, tactile and other sensory perception the key factors can make the participants to produce immersive visual and auditory perception in addition, users can also manipulate virtual objects in virtual objects at the same time, feel the reaction, resulting in tactile and haptic perception. Force perception mainly by the computer through the force feedback glove, force feedback joystick of finger motion damping users can feel the magnitude and direction of force. Tactile feedback is mainly based on the visual sense of touch, pressure, vibration, electronic touch and nerve, muscle simulation and other methods to achieve the main through the basic model. Construction technology, space tracking technology, visual tracking and vision sensor technology, high performance computing technology to achieve.

The accuracy of the virtual environment. The virtual environment that is consistent with the objective world, which requires a wide variety of configurations, complex information to make accurate and complete description. At the same time, need to study the efficient modeling method, reconstruction of virtual object and the evolution rules of all kinds of relationships and interactions.

The virtual environment perception information authenticity synthesis. The abstract information model cannot be directly perceived directly to human, so we need to study the virtual environment of visual, auditory, tactile and haptic synthesis method of perceptual information, focused on solving the problem of high fidelity and real-time synthetic information, in order to improve the sense of immersion.

The nature of interaction between human and virtual environment. The real-time synthesis of perceptual information transfer to the user through the interface, the user according to the sensed information and make analysis and judgment of the events in the virtual environment and situation, and realize the interaction with the virtual environment in a natural way. This requires research based on imprecise information multimodal human-computer interaction pattern and individual natural interaction technology, in order to improve the efficiency of human-computer interaction.

In VR, the computer is from various human movements, such as changes in language information, to correctly understand the information needs to use AI technology to solve, such as speech recognition, image recognition, natural language understanding, research in the field of the intelligent interface is the basis of VR technology, VR technology is also difficult. In essence, to solve the 6 problems mentioned above allows the user to feel personally on the scene of virtual environment, so as to explore and understand the objective things. Generally speaking, the research focuses on virtual reality expansion are all around the problem of the 6 groups.

III. THE REALIZATION FUNCTION AND DESIGN RULE OF HUMAN-COMPUTER INTERACTION INTERFACE

Human-computer interaction is a study of the interactive relationship between the system and users. The learning system can be a variety of machines, can also be a system of computer software and the man-machine interface. Usually refers to the visible part of the user. The user through the man-machine interface and communication system, and operate as small as the radio play button. To the dashboard, aircraft, or control room in power plant. The design of man-machine interface to include the understanding of users of the system, it is for system availability or user friendliness.

At present, human-computer interaction is still exist many problems, mainly from the following three aspects: the use of limited range, has yet to get rid of interactive interface, information is difficult to identify. On the whole, the man-machine interactive way with networking upgrading and the development of artificial intelligence and continuously in the following three aspects: the development of user centric, biometric personalized, full range of perception. The future, communication between man and machine, will be from the mechanical interaction up to the emotional aspects of external communication, information infrastructure all specific operational equipment will naturally melt into the whole, will be replaced by a variety of sensors, everywhere. Various shapes, as well as the integration of artificial intelligence, big data cloud computing platform, they will become more and more intelligent, considerate, real-time human-computer interaction system to provide the Everfount Information.

The traditional way of human-computer interaction, both the interaction between human and computer is through the keyboard, mouse, screen and other tools to achieve. Virtual reality is seen as a unified computing science processing object a computer generated space, and the operation of its people as a part of the space.

The interaction between man and computer space is perceived through a variety of advanced technology and display technology. People can feel the objects in the virtual environment, virtual environment can feel all kinds of operation on it. Virtual environment is man-made, is present in the computer. The user can enter "" this virtual environment, can interact in a natural way and the environment. The so-called interaction refers to the perception of the environment and intervention environment, can let the user generated in the corresponding real environment in the illusory sense of immersion, that is, personally on the scene feeling. And the virtual environment system includes man-machine interface and operator. Computer.

Virtual reality is the main method is with the necessary equipment, to achieve the information conversion between human and virtual environment, achieve natural interaction and interaction between people and environment. And the function of human-computer interaction technology determines corresponding system friendly operation, with the development of technology command more and more complex, the requirements of human-computer interaction more and more high. The history of the development of human-computer interaction is accompanied by the development of.PC software platform, and products bound the body, eyes tied the hands and eyes, which means that the need for a new interactive media as the carrier, VR and AR may be the next generation of media.

In order to maximize impact and achieve the purpose, as in the conception of product positioning should be guided by the following principles: integrity, aesthetic consistency, direct manipulation, feedback, metaphor, user controllable. In the whole system, the text is legible in each size in the figure is accurate and clear. The decoration is delicate and appropriate, this is in order to guide the design of more focus on the function above. Blank, colors, fonts, graphics and interface elements ingeniously highlight the important content and effective interaction.

Sliding gestures and clear nice interface allows people to better interact and understand the content, but not overwhelming. The main contents are usually fill the screen, use more translucent and blur effect to foil. In order to ensure the light and transparent interface must use less borders, gradients and shadows. This will ensure that the content is the highest in the show the importance. The unique visual hierarchy and convey the real action level, give vitality, easy to understand. The touch, the new sense of joy, the use of new functions and new content without distortion. These changes will provide more interesting when you browse the contents of the sense of hierarchy.

IV. Design and Application of Virtual Reality Technology in Human-Computer Interaction Interface of Industrial Products

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Abbreviations and Acronyms

Virtual reality has been some of the world's large enterprises widely applied to all aspects of the industry, the development of the enterprises to improve efficiency, strengthen data collection, analysis, processing capacity, reduce mistakes in decision-making, played an important role in lowering the risk of enterprise. The virtual reality technology is introduced, will make the industrial design method and the thought quality leap, more in line with the needs of social development, can be said to be feasible and necessary in the application of virtual reality technology in industrial design.

Virtual reality is the use of computer simulation to generate a three-dimensional virtual world, to provide users a visual, auditory, tactile and other sensory simulation, such as allowing users to personally on the scene move, there is no limit to observe in three dimensional space. The application of this technology can greatly enhance the human-computer interaction technology, can greatly enhance the user's experience, to further enhance the competitiveness of products.

The world industry has undergone tremendous changes, large-scale sea tactics no longer meet the development of the industry, the application of advanced science and technology shows great power, especially the application of virtual reality technology is a hitherto unknown industrial revolution. Virtual reality has been some of the world's large enterprises are widely applied to various aspects of industrial the enterprise to improve development efficiency, strengthen data collection, analysis, processing capacity, reduce mistakes in decision-making, played an important role in lowering the risk of enterprise. The introduction of virtual reality technology will make the industrial design method and a qualitative leap in thinking, more in line with the needs of social development, can be said to be feasible and necessary the application of virtual reality technology in industrial design.

The industrial simulation system is not a simple scene roaming, is used to guide the production of the simulation system in real sense, it combines the user service layer and database data to set up a complete simulation system can be set up B/S, C/S two kinds of application architecture, and enterprise ERP, seamless, MIS support SqlServer, Oracle, MySql as the mainstream database.

The scope of industrial simulation covers a very wide, from the simple mechanical assembly of single workstation to multiplayer online collaborative training system. Virtual reality provides a new mode to carry out emergency drills, the scene of the accident simulation to virtual scenes, where the manufacturing accidents artificially, make the right in response to the participating organizations staff. This deduction greatly reduces the cost, improves the deduction of training time to ensure that people face accident disaster coping skills, and can break the limits of space and convenient organization of the personnel of deduction, such cases have been applied, will be a trend in the future. Because of emergency exercise it has a simulation, pertinence, openness, autonomy, security features, construct a set of digital open digital resources for the enterprise, through the virtual space inside. When recording, construct a set of emergency drills in the library, and virtual digital environment reproduction corresponding emergency drills, improve their professional level in the virtual environment.

At present, the domestic design and manufacture of industrial products has been using computer aided design, mold making grass rapid prototyping technology, but is still the traditional product design process, the influence of efficiency of product development. Using this technique, the virtual test in product design do some feasibility in industrial design, the specific design operation based on virtual reality in the process of product design to simplify the product, product reviews can be done in real-time dynamic evaluation and virtual three-dimensional products using this technology, can observe any angle of internal and external product structure, product promotion stage, virtual reality technology can directly provide users with a variety of virtual scenes for customers to understand the morphology and visual the effect under different conditions of the product, and allows customers to experience more interactive simulation of the product with an external device, further stimulate consumer The desire to buy and upgrade the product sense of technology and competitiveness.

The design of pre feasibility test, study the needs of users, the feasibility of rapid definition of previous research using virtual reality technology, and presents a variety of forms. The sketch after local kinetic energy rapid and simple characteristics of the product, in a virtual way show, allows ID designers, product planning more clear options more clearly and modify the program.

The ID designer modeling stage, after the completion of the corresponding model can be imported into the software, do product animation design according to the characteristics and functions of products, to provide support for the later product form of display, animation display can bring different visual feelings and stimulate consumption for the user, to enhance the value of product positioning.

Product review is a creative design and production of products can be recognized by the designer, the traditional product review must do product prototype, prototype production time and cost of manpower cost is high, the kind of product prototype for the product design more intuitive, in the review to check the appearance or function model. The rationality of the structure of virtual technology compared with hand only the defect is non physical, but also all the details view of real-time display of products, but also some details of instant change of product in the review, can use virtual technology to create a virtual space and real-time products together, to set the product according to the user's habits and character space. Product review can also create a better the user experience of virtual reality technology and physical space by hand combined. Virtual reality technology for product review more display functions, virtual body Inspection, real-time adjustment of products, products exaggerated visual experience and more interesting review.

International Conference of new products, the virtual reality technology can perfectly play the leading role on the conference, the establishment of a virtual scene in the conference to enable publishers and product integration, perfect display of products in various parts of the details and outstanding function, at the same time also shows a highly interactive. The establishment of virtual product experience center, for excellent products tailored virtual space more creative. Virtual reality display design display used in commercial products. Commercial products display the purpose is to attract people's attention by showing good design, and with the help of multimedia or network means to convey the display information.

Virtual reality interactive technology can we touch things virtual objects into, you can push, grab, and even squeeze them. It can capture the real objects in each space, virtual clone to achieve high quality by 3D modeling, a user can touch things, shorten the people of important life, locations and activities the distance by using this technology, we can achieve more amazing experience in the virtual world.

Virtual reality display design provides such a display means, and the development of computer hardware and network and computer efficient 3D operation to enhance the ability of the virtual reality technology in the display of goods has become even more widespread. The application of multimedia display previous commodity display is the way to the plane pictures, text, visitors can only get by text, voice, graphics, animation and other means of describing single commodity information and virtual reality display. The plane design can break through the limit, is not limited to two-dimensional space of the text, picture and image display methods, which well solves the commodity The lack of realistic problems in the show: on the one hand, 3D virtual display made by virtual reality technology, can let the consumer goods and form a good interaction, make it easier for consumers to grasp the commodity information in order to assess and promote consumer goods, make a purchase decision; on the other hand, instead of using virtual products product display, but also reduces the product cost, from each link to improve the display efficiency.

V. THE FUTURE TRENDS AND DIRECTIONS OF VIRTUAL REALITY AND HUMAN COMPUTER INTERACTION

With the development of virtual reality technology in city planning, military and other aspects of the deep application of interaction in modeling and rendering method, construction method and system of virtual reality technology are put forward higher requirements. In order to meet these new demands, in recent years, virtual reality technology research follow the "low cost high performance" principle, has achieved rapid development, showing some new features and trends. The combination of virtual reality and network communication and multimedia technology, the traditional information technology is a breakthrough in the development of new technology has far-reaching potential applications. Online interactive virtual world is the direction of development of information society, is the inevitable goal of various industries. The future information network interactive virtual reality technology will change people's way of thinking, change the people of the world, their views of space and time.

Throughout the course of development of VR, the future research of VR technology will continue the "low cost, high performance" principle, from two aspects of hardware and software, the main development direction is summarized as follows:

Dynamic environment modeling technology. The establishment of virtual environment is the core content of VR technology, dynamic environment modeling technology to 3D data acquisition environment, and according to the need to build a virtual environment model.

Real time 3D graphics generation and display technology. 3D graphics generation technology is relatively mature, but the key is how to generate real-time, without reducing the quality of graphics and complexity, how to improve the refresh frequency is an important research content in the future. In addition, the development of VR also depends on the stereoscopic display and sensor technology virtual, existing equipment can not meet the needs of the system, it is necessary to develop a new generation of 3D graphics and display technology.

Development of a new type of interactive device. Virtual reality and virtual world objects people can freely interact with input and output devices like personally on the scene, the main data gloves, helmet display, data clothes, 3D position sensor and three-dimensional sound generator.

Intelligent voice virtual reality modeling. Virtual reality modeling is a more complex process that requires a lot of time and effort. If VR technology and intelligent technology, combined with the speech recognition technology, can solve this problem. We model attributes, methods and general characteristics of the description into the necessary modeling the data through voice recognition technology, and then use the computer graphics technology and artificial intelligence technology to design, navigation and evaluation, the model is expressed by the object, and the basic model of static or dynamic connection, and ultimately the formation of the system model.

The prospect of distributed virtual reality technology. Distributed virtual reality is an important direction for future development of virtual reality technology. With the emergence of many DVE development tools and system application, DVE itself has penetrated into all walks of life, including medical, engineering, training and teaching and collaborative design. Simulation training and teaching training is another important application areas DVE, including the virtual battlefield, assisted teaching.

VI. CONCLUSION

As a comprehensive reflection of the frontier of modern science and technology, VR art is a new art form of language visualization and interaction of complex data through man-machine interface, it is important to attract artists, in close combination of artistic thinking and technology tools and two deep penetration of the new cognitive experience. Compared with the traditional the windows operating under the new media art, interactive and extended dialogue is the key to VR art has its unique advantages. From the overall sense, VR is a new man-machine interactive art based art form, its biggest advantage is that the construction works and participants of the dialogue, through dialogue and reveal the significance of the process generation.

In combination with the existing work mode and work content, the introduction of virtual reality technology will be one of the highlights of the industrial design, the SKYWORTH brand will also enhance the sense of science and technology, further reflect the focus on health technology. Virtual reality is the future of human-computer interaction, product design, display the development trend.

The application of virtual reality technology, the product design must take an important role, not only on the product design, should also actively explore some new areas, so that the virtual reality technology in industrial design involves more areas. Science and technology development today, virtual reality technology is constantly improving. The continuous development, continue to play a greater value in the design, produce the maximum energy, the majority of users to the most perfect products and most people enjoy the visual perception and experience.

REFERENCES

[1] Xu Shouxiang, Hu Wen, Jackie Chan, Ma Chao. Multi channel virtual reality interactive terminal design and its application [J]. Journal of Shenzhen Institute of Information Technology, 2015, (03): 22-26.

[2] Wang Yu. China's animation of the mobile phone communication studies [D]. Northeast Normal University, 2013.

[3] to soar. Interactive display and interaction design for the grand view, 2013 cases of [J]. art application of virtual reality in the field of art and design, (03): 100.

[4] Zhang Lu. Virtual reality technology, user interface design and research based on [D]. of Donghua University, 2013.

[5] Chen Zhigang. The 3D digital product prototype interactive display design of [D]. Jiangnan University, 2010.

[6] Wang Guowei. User centered design of [D]. virtual interactive roaming system of Zhejiang University, 2010.

[7] Hu Yan. The interactive virtual reality system research on the key technology of [D]. Shaanxi Normal University, 2009.

[8] Li Chunfu, Li Zexiang. Virtual reality in the industrial application in the design of [A]. State Intellectual Property Office Design Review Department,.Proceedingsofthe2008InternationalConferenceonIndustrialDesign Chinese Institute of mechanical engineering industrial design branch of the State Intellectual Property Office (Volume1) [C]. design review department, Chinese Institute of mechanical engineering, industrial design branch: 2008:4.

[9] Ren Jie, sun Surong. Virtual reality technology application status and development trend in industrial design in [A]. China Institute of mechanical engineering industrial design

branch.Proceedingsofthe2007InternationalConferenceonIndustrialDesign (Volume1/2) [C]. China Institute of mechanical engineering industrial design branch: 2007:5.

[10] Jiang Zibin. Research on [D]. interactive design of 3D landscape of Central South University, 2003.

[11] Project number: of Anhui Province, humanities and Social Science Key Project: project name, application of virtual 3D animation technology in product design in human-computer interaction interface (SK2014A0673)The research project of Anhui Xinhua University research team.

[12] Project number: animation application of Huizhou culture in the animation creation are: project name, application of virtual 3D animation technology in product design in human-computer interaction interface (2016td004)

The Study of Following Behavior to Bi-direction Pedestrian Flow with the Dynamic Preconscious Effect

Xin Tang
School of Logistics Engineering, Wuhan University of Technology, Wuhan 430063, China

Lin Pan, Jiying Wang
School of Logistics Engineering, Wuhan University of Technology, Wuhan 430063, China
e-mail: linpandr@163.com;
wangjiyingyichang@163.com

Xueyu Zhao
National Engineering Research Center for Water Transport Safety, Wuhan University of Technology, Wuhan 430063, China
e-mail: stzxy@whut.edu.cn

Yi Yang
School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430063, China

***Abstract*—In view of the preconscious behavior of pedestrian and walking speed differences, a lattice gas model of bi-direction pedestrian flow is established in this paper. According to the characteristics of pedestrian following behavior and preconscious dynamic change in different walking conditions, bi-direction pedestrian behavior model based on dynamic preconsciousness is constructed to study the bias decision-making behavior of pedestrian movement. Through numerical simulation, the influence of regional size, asymmetry and speed deviation on the Bi-direction pedestrian flow is analyzed. Results indicate that dynamic preconscious behavior enhances the stability of the system and reduces pedestrian congestion. The passing behavior of the high-speed pedestrian is the main cause for the congestion in the situation of high density. The speed difference will largely influence the anti-congestion ability of the system, so keeping the unity of the speed and group order would maintain the stability and the anti-congestion ability of the system for the whole system.**

Keywords-bi-direction pedestrian; dynamic preconscious effect; following behavior; lattice gas model; probability distribution

I. INTRODUCTION

In recent years, there has been a frequent occurrence of emergencies due to the entire channel system congestion caused by bi-direction pedestrian conflict. This phenomenon has resulted in greater property damage and a threat to the physical safety of pedestrian traffic participants. Thus, the effective control of pedestrian congestion is of extreme importance. As the main component of the walking traffic system, pedestrian individuals can produce strong social force between the characteristics with autonomous behavior, a greater degree of random decision-making, compressed space, and the non-fixed shape. Therefore, pedestrian research is often more difficult than motor vehicle traffic. As a result, it is of great significance to establish a reasonable pedestrian flow model to explore its macro-behavior characteristics and the formation of congestion mechanism, which contributes to enhance the level of basic theory and application on pedestrian traffic in China.

In 1971, Henderson presented a macroscopic model of the pedestrian flow, arguing that pedestrians in free-flowing states had properties that were similar to those of gas molecules. This model can describe the macro characteristics of the traffic flow in detail, but the micro characteristics of the traffic flow cannot be described. Therefore, there is no scientific explanation for self-organizing phenomena in the traffic flow [1]. In 2002, Hughes improved the pedestrian flow model and proposed the dynamic structure of macro pedestrian flow, regarding the pedestrian behavior as a continuous fluid described by two-dimensional spatial density evolution mechanism [2]. Thereafter, Hoogendoorn and Bovy proposed a deterministic pedestrian equilibrium distribution model based on the assumption of complete traffic information [3,4]. Through the previous study, Huang further put forward the condition of the pedestrian equilibrium condition and realized the numerical solution, which verified the model scientifically [5,6]. Based on the ability of the pedestrian to respond to the route and the environmental memory, Xia proposed a hybrid selection strategy to explore the bypass problem of pedestrian [7, 8]. Jiang reflected the congestion characteristics of the traffic flow through the numerical simulation of the high order macro flow model [9]. By using the relationship between rate and time to establish the pedestrian flow behavior model of single row longitudinal channel, Lv et al. carried out the pedestrian movement evaluation and evacuation model of public export channel [10]. Hoogendoorn et al., for example, consider the behavior of path selection before and after the departure, and proposes the pedestrian flow model of multi-level continuous medium [11]. Hänseler et al. established a macro-load model of time-bound traffic in the public travel area, which can describe the isotropic framework of real-time potential conflict propagation of multiple pedestrian groups [12]. Considering the factors of impact of local density and personal space, safe distance, neighbors, direction and other information together, Xiao et al. introduced the Voronoy diagram method into the pedestrian flow heuristic analysis model [13]. Fu et al. carried out a simulation study of the popular epidemic patterns of emotional transmission and found that the spread of panic sentiment with the trampling incident would lead to an increase in the expected speed of pedestrian flow [14].

In retrospect of the pedestrian flow model, most scholars tend to use the uniform system which is a system that does not consider the variability of pedestrian properties. For instance, pedestrian's own characteristics are ignored and the unlimited speed is set to $V_{\max} = 1$, which is mistaken in reality.

In addition, influenced by regional culture, laws and regulations, pedestrian characteristics and customs, pedestrians unconsciously develop certain preconscious behaviors in the course of their progress. For example, in the country or region whose provisions of the road transport vehicles on the right, such as mainland China or the United States, pedestrians with many years of experience, unconsciously show some behaviors. They unconsciously walk along the right side of the road while exceed from the left, as well as avoiding the opposite pedestrian flow on the right. Meanwhile, preconscious awareness can lead pedestrians to make independent judgments about the surrounding traffic and even make decision predictions ahead of time. As a result, compared with the conventional uniform system, in the pedestrian flow modeling, if the pedestrian's previous consciousness caused by the movement trajectory and the rate of different can be shown more, the reality of pedestrian flow situation can be more truly reflected.

Preconsciousness, the intermediate link between the subconsciousness and consciousness, is the awareness that people can predict the occurrence and consequences of others or their own affairs in advance. Unlike the subconscious instincts, the preconsciousness reflects more of the behavioral patterns of pedestrians making quick decisions and judgments about the form of traffic. Also, this behavior pattern may be dynamically adjusted depending on the form of change. This behavior pattern not only stems from unconscious thinking and years of experience, but also reflects the pedestrian's independent judgement on the surrounding traffic environment, even the psychological characteristics of early decision making. And this is an indispensable part of the action decision. Factors that affect the preconscious behavior of pedestrians include regional culture, rule of law, pedestrian characteristics and customs. In the modeling of pedestrian flow, If the difference between movement trail and speed caused by pedestrian's preconsciousness is more shown, the authentic pedestrian flow would be more reflected.

Aiming to simplify the analysis, we do not consider the factors that affect the speed of the pedestrian, and simply put forward the hypothesis that the pedestrian is classified as high-speed and low speed. In another words, only two types of pedestrian speed are set. To avoid the collision and interference with the opposite traffic flow, pedestrians tend to develop the habit of walking on one side of the road, which is also related to the country's traffic rules. For instance, people in mainland China generally show a right-walking habit, while the British or Japanese do the opposite [15,16]. Based on these considerations, this paper mainly studies the preconsciousness of right-side walking. At the same time, pedestrians feel a strong sense of security due to going along the right side of the road. Their nerves relax and they lower the speed unconsciously. Hence, under the reality circumstances, low-speed pedestrians tend to move along the right side of the road, which also reflects the differences in the specific performance of pedestrians. High-speed pedestrian due to preconsciousness will tend to exceed the low-speed pedestrian of the same direction from the left side. In the process of opposite pedestrian interweave, pedestrians tend to move on the right side to avoid collision and conflict. In general, pedestrians have preconscious behaviors that tend to walk on the right side, exceed from the left, and avoid to the right side. While in practical pedestrian traffic, pedestrians tend to walk safely and comfortably, following the same pedestrian movement in order to reduce unnecessary conflict. In the process of movement, according to their surrounding environment and the surrounding circumstances, pedestrians decide whether to accelerate, slow down, follow or change their movements. Different local environments cause pedestrians to follow behavior in different enclosed spaces. And the magnitude of waiting probability is changing. For example, when a pedestrian finds that there is a walker in front of him moving at the same speed, he will be more inclined to trail after. In this case, the intensity of the behavior in the enclosed space, which is the waiting probability, is relatively large. However, the current pedestrian will have the preconscious behavior to exceed the person from the left direction who is walking at a low speed in front of him. At this time, the waiting probability is relatively slight. As a result, under different circumstances, studying the changes of pedestrian following intensity will contribute to further depict the micro-motion of a pedestrian. The characteristics of macroscopic behavior of traffic flow and the formation of complex phenomena would be thoroughly understood.

II. MODELING CONSTRUCTION AND ANALYSIS OF OPPOSITE PEDESTRIAN'S TRAILING BEHAVIOR BASED ON DYNAMIC PRECONSCIOUSNESS

A. *Modeling Analysis of Pedestrian's Trailing Behavior in Enclosed Space Based on Preconscious Behavior*

As shown in Fig.1, pedestrian's modeling in enclosed space is established in a two-dimensional system bounded above and below. The length of the system is L and the width is W. the pedestrians are prohibited to cross. There are four types of pedestrians defined in the system: high-speed pedestrians to the right (right triangle), low-speed pedestrians to the left (circle), high-speed pedestrians to the left (left triangle) and low-speed pedestrian to the left (cross). The pedestrian can only occupy one frame and cannot move backward.

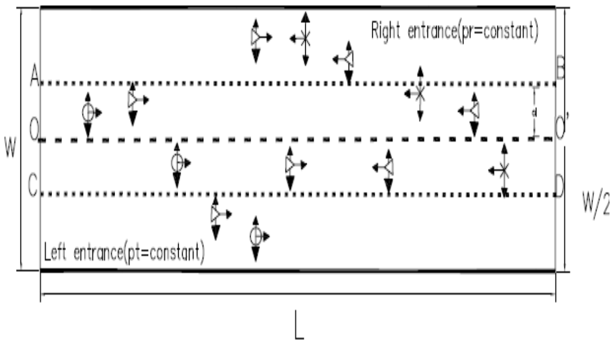


Figure 1. The multi-type pedestrian system in the enclosed space

The arrows indicate the direction that pedestrians can choose, and the bold arrow indicates the direction of motion that makes the person prone to preconsciousness. In another words, different types of pedestrians in the process will be subconsciously move to the area that they feel safe and comfortable. For example, when high-speed and low-speed pedestrians to the right are located in the upper parts of the middle line OO', they all have the right side preconsciousness in the movement. That is to say that they tend to move towards the lower part of the middle line OO' and the low-speed pedestrian to the right are more likely to walk in the lower area of the cutting line CD.

Therefore, based on real circumstances, the stable speed of high-speed (the average speed of free flow which is regarded as the maximum speed) pedestrian's movement is about 1.6 m/s, the stable speed of low-speed pedestrian is about 0.8 m/s. Then the actual time for each step in the simulation is about 0.5s.

Given the respective special space state in preconsciousness following behavior of high-speed and low-speed pedestrians, it is necessary to set a low-speed and high-speed to right direction in the space that the pedestrian may encounter during the course of the movement. As shown in Fig.2 and Fig.3, the blocking states that pedestrians may encounter from front, left and right direction are presented. The solid circles refer to the pedestrians to the right direction. Bold arrows indicate the possible direction of motion. The crossing shapes are on behalf of the position occupied by other pedestrians, and the corresponding movement directions are indicated by the thin arrows. In the system, pedestrians can only choose to follow the arrows or wait.

The four probabilities shown in the Fig.2 and Fig.3 are explained as follows: Pt,x refers to the probability of moving to the right direction, Pt,y stands for the probability of moving up (moving to the right direction), Pt,-y means the probabilities of going down (moving to the right direction), Pw represents the probability of waiting. After thinking about the preconsciousness of pedestrians, based on the classic lattice gas model, some theoretical formulas have been improved. Except that the original reference D indicating the bias intensity of heading direction, reference B is newly added to identify the bias angle of preconscious direction.

To truly reflect the complex behavior of pedestrians, the different sets of probabilities of moving to right direction should meet the following rules.

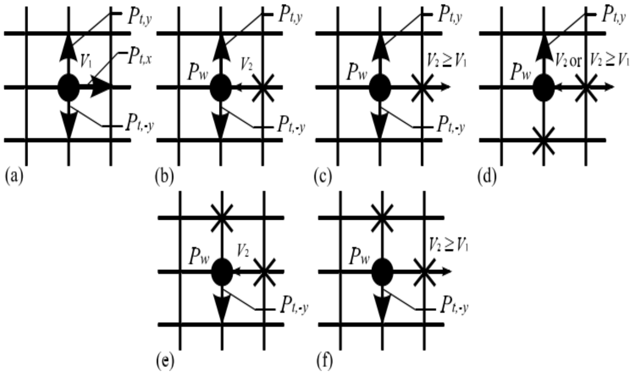


Figure 2. The space state and walking probability distribution of the right-direction pedestrians with low speed

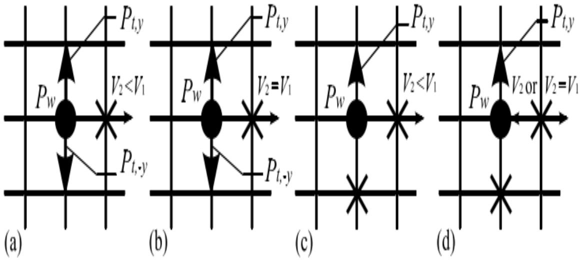


Figure 3. The space state and walking probability distribution of the right-direction pedestrians with high speed

As a low-speed pedestrian moving toward right direction, if the front lattice point is not occupied, generally low-speed pedestrians do not choose to wait. However, there exists some probabilities of moving toward left or right direction, which are related to bias intensity of moving forward and preconscious direction (left or right side). If the front lattice point is occupied by right pedestrian in the same direction while the right lattice point is not occupied, no matter how fast if the pedestrian, the waiting probability of the pedestrian is higer. Because of under this circumstance, pedestrians are more willing to keep their direction of motion and follow the pedestrians of the same direction. If the front lattice point is occupied by the pedestrians of the opposite direction, then the probability pt,-y of moving to the right direction is high. The reason is that pedestrians have a preconscious bias habit towards avoiding collisions to the right direction. if the right and front lattice grid are both occupied by other pedestrians, then the waiting probability Pw of this pedestrian is high. It is because that when the left-side pedestrian is ahead, the low-speed right-side pedestrians following the right-side traffic rules and preconsciousness are inclined to keep stationary, expecting to avoid collision with the left-side pedestrian moving toward right direction. As shown in Fig.2(d), when the right-side pedestrian is ahead, the following behavior in enclosed space will be more likely to occur.

As a high-speed pedestrian moving toward right direction, if the front lattice point is not occupied, then the behavior is the same as the low-speed pedestrian moving toward right direction, shown in Fig. 3(a). If the front lattice point is occupied by the low-speed pedestrian in the same direction, then the probability Pt,y of moving toward left direction is low. The reason is that pedestrians develop a preconscious habits exceeding from the left side, which is shown in Fig. 3(a) and (c). If the front lattice point is occupied by the high-speed right-side or opposite-side pedestrians, then the probability Pw is high. This is the result of following effect and right-side bias preconsciousness, which is shown in Fig. 3(b) and (d).

To show the speed difference between the different types of pedestrians, the sports rules of FI cellular automation model is introduced. If pedestrians meet the requirements of walking probability Pt,x, then in unit time (time step taken as the unit), these pedestrians can move forward x(n) lattice points. The value of x(n) is the small value of the nth pedestrian’s maximum speed and the number of spaces. Because of human nature's great flexibility and adaptability, the individual can realize short time acceleration. Stable velocity can be achieved without obstacles ahead. Therefore, it is reasonable to choose the sports rules of FI model.

Based on these analyses, in view of these two different pedestrian space conditions, the space state distribution of low-speed pedestrians is as follows:

$$P_L=\left[p_{t,x},p_{t,y},p_{t,-y},p_w\right]=\begin{bmatrix} D+\frac{1-B}{3} & \frac{(1-B)(1-D)}{3} & \frac{(1+B)(1-D)}{3} & 0 \\ 0 & \frac{(1+B)(1-D)}{3} & D+\frac{1-D}{3} & \frac{(1-B)(1-D)}{3} \\ 0 & \frac{(1-B)(1-D)}{3} & \frac{(1+B)(1-D)}{3} & D+\frac{1-D}{3} \\ 0 & \frac{(1+B)(1-D)}{2} & 0 & D+\frac{(1-B)(1-D)}{2} \\ 0 & 0 & D+\frac{(1-B)(1-D)}{2} & \frac{(1+B)(1-D)}{2} \\ 0 & 0 & \frac{(1-B)(1-D)}{2} & D+\frac{(1-B)(1-D)}{2} \end{bmatrix} \quad (1)$$

The space state distribution of high-speed pedestrians:

$$P_L=\left[p_{t,x},p_{t,y},p_{t,-y},p_w\right]=\begin{bmatrix} 0 & D+\frac{1-D}{3} & \frac{(1-B)(1-D)}{3} & \frac{(1+B)(1-D)}{3} \\ 0 & \frac{(1+B)(1-D)}{3} & \frac{(1-B)(1-D)}{3} & D+\frac{1-D}{3} \\ 0 & D+\frac{(1+B)(1-D)}{2} & 0 & \frac{(1-B)(1-D)}{2} \\ 0 & \frac{(1-B)(1-D)}{2} & 0 & D+\frac{(1+B)(1-D)}{2} \end{bmatrix} \quad (2)$$

Similarly, we can see that the preconscious behavior of pedestrian characteristics to the left and right direction is exactly the same.

B. *Dynamic Preconscious Intensity and Behavior Modification Based on Early Warning Perceived Distance*

Preconscious intensity discussed in the previous section is set to a static parameter. In the reality, however, pedestrians are able to adjust the direction of the trend according to the position and characteristics of their environment when walking. As a result, there is certain dynamism in the bias preconscious intensity of pedestrian. This dynamic characteristic is reflected in the condition that a pedestrian is constantly correcting his or her distance from the pedestrian or obstacle in front of him to choose the opportunity to avoid or exceed. At the moment, the bias preconscious intensity increases with the distance of the person ahead. The probability of a pedestrian being able to exceed or avoid in the same direction in advance is also increasing. As a result, without considering the specific factors of pedestrian perception, the forward traveler early warning perceived distance parameter and self-admissible distance is added in the paper. Then the parameter B can be approximated as:

$$B=B_0(1+\frac{D_2-D_1}{D_2}) \quad (3)$$

Formula (3) stands for initial preconscious intensity. That means pedestrians feel comfortable and safe before seeing pedestrians ahead. With pedestrians ahead entering the awareness range, people will feel sense of urgency and the bias preconscious intensity will be increased to boost the deviation probability. The parameters change depending on the person's personality. Generally speaking, the distance that casual and careless people can mentally endure is always big. While for those who are sensitive, the distance will be smaller. In addition, due to the limitation of increased consciousness intensity, a threshold value can be set for intensity. When the conscious intensity of pedestrians has achieved to the maximum value, it will not grow as the pedestrian approaches. When a pedestrian perceives that someone in front of him will enter his or her perceived warning distance, the distance between them actually has to do with the difference value of velocity vector. When the two walk oppositely to each other, the reduction rate of D1 is the sum of the two’s speed. However, when the speed of the pedestrian in the front is smaller, the reduction rate of one is the difference in the quantity of the two. In conclusion, after introducing the warning timing variable (The step is calculated from the forward warning distance), the formula (3) can be rewritten as:

$$B=\min\left[B_{\max},B_0(1+\frac{(V_i-V_{i+1})\Delta t}{D_2})\right] \quad (4)$$

After introducing the dynamic preconsciousness, for the people who enter the self-admissible distance in the same or opposite direction, pedestrians may respond with preconsciousnesss in advance. In another words, pedestrians may take the bias or waiting action in advance before achieving the people ahead (the front lattice point has not been occupied by other pedestrians). Therefore, the range of lattice points occupied by pedestrians ahead can be extended from one lattice point to all lattice points within perceived warning distance, which can be calculated by formula (1) and (2).

III. NUMERICAL SIMULATION AND ANALYSIS

In view of enclosed space condition, numerical simulation and analysis of the model are carried out in this paper. At the beginning of system operation, four types of pedestrians are randomly distributed within the system. The values are generated by random sequences and compared with direction converted probability to implement the location update for the pedestrian. Meanwhile, the system is set as a periodic boundary. When the pedestrian to the right direction reaches the right boundary and disappears, from the left side the regenerated individual enters the system. And for the pedestrians to the left direction, the same method is set [14]. As a result, in every time step, the number of people in the system is always kept constantly.

Parameter m and h are introduced in this paper to represent the ratio of pedestrian to the left and high-speed pedestrian respectively. The non-equilibrium characteristics of system pedestrian behavior are shown.

On this basis, the total density of the important representational parameters, average speed and average flow rate of the simulation result is set in this paper. The total density is the ratio of total pedestrian number and system area size ($W \times L$). The average speed is defined as the ratio of moving speed and the total pedestrian number in a single time step. The average flow is the number of pedestrians passing through the destination boundary (ie, right-hand pedestrians through the right border, or left-hand pedestrians through the left border). According to the measurement needs, this paper takes the mean flow from the right flow and the leftward flow as the average flow. To sum up, the parameters are expressed as (5) - (7).

$$p = \frac{N}{W \times L} \tag{5}$$

$$\overline{V} = \frac{1}{TN} \sum_{t=t_0}^{T+t_0-1} \sum_{i=1}^{N} v_i \tag{6}$$

$$\overline{V} = \frac{1}{TN} \sum_{t=t_0}^{T+t_0-1} \sum_{i=1}^{N} v_i \tag{7}$$

The numerical simulations would use an average of 20 samples to reduce the effect of initial random distribution effect. Every random sample runs 10,000 time steps. The average speed and the average flow are to take the last 8000 steps to calculate the results.

The relationship curve between average speed and average flow and pedestrian density in different area sizes is shown in Fig.4. In Fig.4 (a) (c), the size of W is changed successively. In Fig.4 (b) (d), the size of L is changed in turn.

As shown in Fig.4 (a) (c), when the pedestrian density is less than critical density, as W grows, the average velocity decreases, but the flow is increasing. The reason is that in the same density, the increase of W enlarges the area of space, so the number of pedestrians will increase, resulting in increased interference between pedestrians. Pedestrians get more resistance and can't move smoothly, so the average speed drops. Meanwhile, as shown in Fig.4 (a), when W is over 40, the system's congestion density is approximately 0.74. It indicates that when L is the same and W is larger, congestion density is no longer affected by region size. Besides, as shown in Fig.4 (a) a system space with a square (width of 100), the transition of the pedestrian flow is more gradual than the other types. In addition, comparing all the lines, we can know that the closer the W is to the L, the transition process of traffic congestion will be slow.

From Fig.4 (b) (d), when the pedestrian density is less than critical density, for different L, the average velocity is the same as the average flow rate, and the variation trend is consistent. However, when pedestrian density exceeds critical density, average velocity and average flow rate decrease rapidly with density. Then the trend slowed until it was completely congested. The increase of L will accelerate the transition from free flow to congestion. In addition, when L exceeds 150, the system's congestion density is approximately 0.72 constantly, which indicates that in the same area, the congestion density is not affected by region size when L is large. In conclusion, in the same regional condition, increasing the ratio of the width of the system to the length of the system will improve the system's ability to resist congestion.

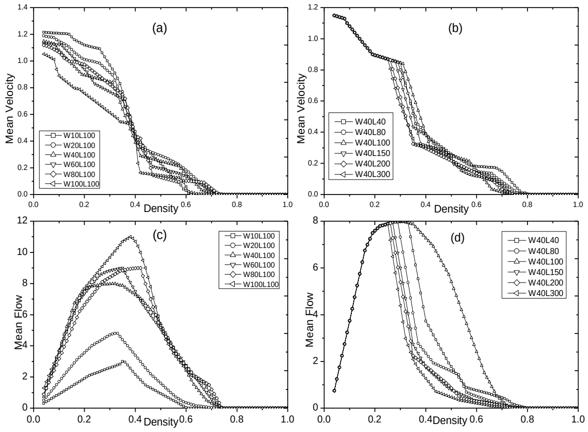


Figure 4. The flow diagram of average speed and flow as density changes with different size of area

Fig.5 shows a diagram of the relationship between the average velocity and average flow over density when m is different. As shown in Fig.5, the congestion density gradually decreases as the m changes from 0 to 0.5. When m = 0.5, the congestion density is minimal.

This is because the pedestrian crowd can segregate from the organization when there is a larger difference in the direction, which can avoid confrontation between pedestrians leading to the decreases between pedestrian’s interactions. Then the flow efficiency of the whole system has been improved. Therefore, the non-equilibrium in the number of pedestrians can enhance the stability of two-way pedestrian flow and improve the system's anti-congestion ability. When the difference between the numbers of pedestrians is small, the system is more prone to congestion. Consequently, the proportion of left and right pedestrians can be controlled rationally according to the crowding level of the pedestrian to optimize pedestrian traffic.

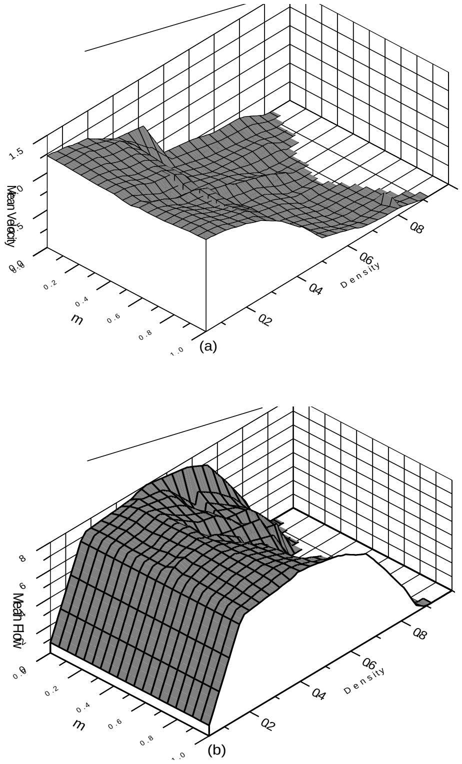


Figure 5. The flow diagram of average speed and flow as density changes with different setting of m

Fig.6 shows a diagram of the relationship between the average velocity and average flow over density when h is different. As shown in Fig.6, in the free flow condition, as h increases, the average speed and flow rate increases. But when h = 0, the density of congestion is greater than when h > 0. In this condition, pedestrians keep moving at a low speed. When the pedestrian density is low, the pedestrian will have greater movement space. The speed of them is little lower than the pedestrians ahead in the same direction. Then they can keep a large distance. The preconscious intensity is low and the change is stable. So there are few acts of exceed, bias or stopping. And moving at the maximum speed is adopted. At this point, the average speed and flow rate increases as the number of pedestrian increases unceasingly. As pedestrian density increases, the interaction between pedestrians increases. In addition, as the number of high-speed pedestrians increased, the probability of the difference between pedestrians front and back increased. The intensity of the preconsciousness also fluctuates significantly. As a result, high-speed pedestrians are more inclined to show the preconscious acts of exceeding (or exceeding in advance). And they may enter the moving area of pedestrians in the opposite direction.

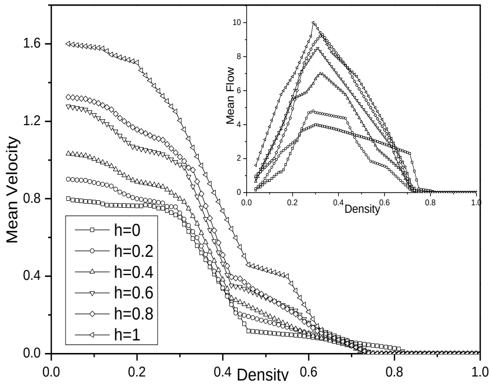


Figure 6. The flow diagram of average speed and flow as density changes with different setting of h

Meanwhile, as the number of high-speed pedestrians increases, it is easier to increase the intensity of the preconscious intensity of the pedestrian in the opposite direction. Then the increase in probability of taking early avoidance behavior (or waiting in advance) leads to a mutually obstructive situation between the pedestrians from the opposite direction. The collision between the pedestrians also intensifies, inducing small congestion and eventually blocking. Based on the above analysis, we believe that behavior of high-speed pedestrians' exceeding preconsciousness in the opposite direction is the major cause of pedestrian congestion. Preconsicousness behavior of avoiding in advance induced by it can aggregate congestion of pedestrian flow.

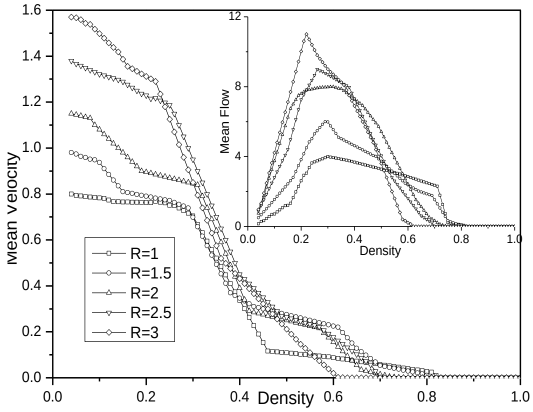


Figure 7. The flow diagram of average speed and flow as density changes with different setting of R

Without changing the ratio of high-speed and low-speed pedestrians, speed ratio parameter R is introduced into this paper. As shown in Fig.7, by adjusting the value of R, average velocity and average flow rate varying rule with density in the condition of different velocity can be further explored. From the Fig.7, we can see that as the ratio of pedestrian rates increases, the density of the congestion decreases, which indicates that the system's ability to resist congestion has gradually declined. At the same time, average speed and flow are also increasing. When R is smaller, the pedestrians are in the condition of free flow. There will be large space between the low-speed pedestrians. The intensity of the pre-consciousness will not fluctuate significantly. Therefore, there is less interaction between the pedestrians. They go their own way and the system is more resistant to congestion. When R is at a high level, the speed gap between pedestrians will increase. The bias preconsciousness of the pedestrian changes faster, whether the pedestrians are in the same or opposite direction. So they are more likely to show the acts of biased or waiting in advance based on awareness. In this case, the conflict between pedestrian and interference behavior occurrence probability is increasing and the system will gradually appear local congestion, until the entirety completely blocked. It is seen that velocity difference between pedestrians will largely affect the ability of resisting congestion. And there is contradiction between the two factors. Keeping the pedestrian speed unity and group order will maintain the stable and anti-congestion ability of the whole traffic system.

IV. CONCLUSIONS

Based on the lattice gas behavior model, we consider the direction probability value in different preconsciousness situation. Multi-speed pedestrians flow lattice gas model in opposite direction varying with intensity has been established. In the simulation part, we study the effects of system size, asymmetry, velocity difference and preconscious behavior intensity on the one-directional and bidirectional pedestrian flow. And the mechanism of the macroscopic phenomena of the pedestrian flow induced by following effects has been discussed. The following conclusions have been obtained.

Preconscious behavior is one of the main reasons for the overall well-organized traffic flow and the resulting self-organizing behavior. Pedestrian lattice gas model of dynamic preconsciousness based on warning distance can show some essential features or phenomena of bidirectional pedestrian flow and following behaviors, such as self-shunt, exceeding and avoiding in advance, waiting and so on. When the system width and length are large, congestion density will no longer change with system size. The ratio of width to length has some effect on the system's ability to resist congestion. The non-symmetry of the number of left and right pedestrians increase the stability of the system and can reduce congestion by self-shunt. Under the high density, exceeding behavior of high-speed pedestrian is the main reason of inducing the congestion of pedestrian flow. The preconscious behavior of bi-direction avoidance in advance will aggregate pedestrian flow congestion.

In conclusion, some research results of this article can provide useful reference for the rational design of buildings and the congestion control of pedestrians. For example, an obvious export symbol (such as “emergency exit” “safety exit ”) or the arrows indicating the direction of escape may be established within the building to avoid blindly following the pedestrian evacuation, thus improving the efficiency of evacuation. Taking advantage of research results of the asymmetrical effect, the ratio of left and right pedestrian number can be controlled properly by the congestion of opposite pedestrians to optimize pedestrian traffic.

ACKNOWLEDGMENT

This work was supported by the Fundamental Research Funds for the Central Universities and ESI Program of China.

REFERENCES

[1] Henderson L.F., The statistics of crowd fluids[J], Nature, 1971, 229(5): 381-383.

[2] Hughes R.L., A continuum theory for the flow of pedestrians[J], Transportation Research Part B, 2002, 36(6): 507-535.

[3] Hoogendoorn S.P. and Bovy P.H.L., Dynamic user-optimal assignment in continuous time and space[J], Transportation Research Part B, 2004, 38(7): 571-592.

[4] Hoogendoorn S.P. and Bovy P.H.L., Pedestrian route-choice and activity scheduling theory and models[J], Transportation Research Part B, 2004, 38(2): 169-190.

[5] Huang L., Xia Y., Wong S., Shu C., Zhang M. and Lam W., Dynamic continuum model for bi-directional pedestrian flows[J], Engineering and Computational Mechanics, 2009,162(3): 67-75.

[6] Huang L., Wong S.C., Zhang M.P., Shu C.W. and Lam W.H.K., Revisiting Hughes' dynamic continuum model for pedestrian flow and the development of an efficient solution algorithm[J], Transportation Research Part B, 2009, 43(1): 127-141.

[7] Xia Y.H., Wong S.C. and Shu C.W., Dynamic continuum pedestrian flow model with memory effect[J], Physical Review E, 2009, 79(6): 066113.

[8] Yanqun Jiang,S.C. Wong,Peng Zhang,Ruxun Liu,Yali Duan,Keechoo Choi. Numerical simulation of a continuum model for bi-directional pedestrian flow[J]. Applied Mathematics and Computation. 2011 (10) :136-157.

[9] Jiang Y.Q., Zhang P., Wong S.C. and Liu R.X., A higher-order macroscopic model for pedestrian flows[J], Physica A, 2010, 389(3): 4623-4635.

[10] Lv W, Fang Z, Wei X, et al. Experiment and Modelling for Pedestrian Following Behavior Using Velocity-headway Relation[J]. Procedia Engineering, 2013,62:525-531.

[11] Hoogendoorn S P, van Wageningen-Kessels F L M, Daamen W, et al. Continuum modelling of pedestrian flows: From microscopic principles to self-organised macroscopic phenomena[J]. Physica A: Statistical Mechanics and its Applications, 2014,416:684-694.

[12] Hänseler F S, Bierlaire M, Farooq B, et al. A macroscopic loading model for time-varying pedestrian flows in public walking areas[J]. Transportation Research Part B: Methodological, 2014,69:60-80.

[13] Xiao Y, Gao Z, Qu Y, et al. A pedestrian flow model considering the impact of local density: Voronoi diagram based heuristics approach[J]. Transportation Research Part C: Emerging Technologies, 2016,68:566-580.

[14] Fu L, Song W, Lv W, et al. Multi-grid simulation of counter flow pedestrian dynamics with emotion propagation[J]. Simulation Modelling Practice and Theory, 2016,60:1-14.

[15] Weng W.G., Chen T., Yuan H.Y. and Fan W.C., Cellular automaton simulation of pedestrian counter flow with different walk velocities[J], Physical Review E, 2006, 74(3):036102.

[16] Yang L.Z., Li J. and Liu S.B., Simulation of pedestrian counter-flow with right-moving preference[J], Physica A, 2008, 387(1): 3281-3289.

[17] Seyfried A., Portz A. and Schadschneider A., Phase coexistence in congested states of pedestrian dynamics[C] .In: Bandini S, et al., editors. Cellular Automata, Springer-Verlag Berlin Heidelberg, 2010, 12(6): 496-505.

[18] Muramatsu H. and Nagatani T., Jamming transition in two-dimensional pedestrian traffic [J], Physica A, 2000, 275(6): 281-291.

[19] Muramatsu M. and Nagatani T., Jamming transition of pedestrian traffic at a crossing with open boundaries [J], Physica A, 2000, 286(5): 377-390.

The Application of Improved PSO Algorithm in the Geometric Constraint Solving

Tian Wei
School of Information Engineering
Changchun Sci-Tech University
Changchun, China
e-mail: 1311895012@qq.com

Zhu Xiaogang
School of Automotive Mechanical Engineering
Changchun Sci-Tech University
Changchun, China
e-mail: 578710782@qq.com

***Abstract*—Geometric constraint solving is a hot topic in the constraint design research field. Particle swarm optimization (PSO) is a method to solve the optimization problem from the biological population's behavior characteristics. PSO is easy to diverge and fall into the local optimum. There are various kinds of improvements. In addition to improving some performance, the corresponding cost is paid. In this paper, a particle swarm optimization algorithm based on the geese is adopted to solve the geometric constraint problem. The algorithm is inspired by the flight characteristics of geese; each particle follows the optimal particle in front of it to keep the diversity; each particle can share more useful information of other particles, which strengthens cooperation and competition between particles. The algorithm balances the contradiction between the search speed and the accuracy of the algorithm to a certain extent. Experimental results show that the proposed algorithm can improve the efficiency and convergence of geometric constraint solving.**

Keywords-PSO; Geometric constraint solving; geese; individual extreme; global extreme

I. INTRODUCTION

With the development of computer technology, CAD/CAM has been developed rapidly. The development and application level of CAD/CAM has become an important sign of the national modernization level. Geometric constraint solving is a hot topic in the constraint design research field. A constraint describes a contented relationship. If users have defined a series of relationships, the system will automatically choose the appropriate state to satisfy the constraints after the parameters are modified. This method is called constraint model. Now many scholars study deeply the constraint solving by using numerical calculation theory, artificial intelligence theory, graph theory, freedom analysis theory. There are the integrated solution, the sparse matrix, connection analysis, protocol construction, constraint propagation, symbolic algebra and auxiliary line [1].

Particle swarm optimization (PSO) is a method to solve the optimization problem from the biological population's behavior characteristics. PSO is easy to diverge and fall into the local optimum. There are various kinds of improvements. In addition to improving some performance [2], the corresponding cost is paid. In view of the above shorts, according to the flight characteristics of geese, the paper proposed two improvements: firstly, global extreme value is transformed into individual extreme of the anterior superior particle value according to historical optimum fitness sorting. So all particles do not direct the same solution, which can avoid the same, maintain diversity, and expand the search scope; secondly, each particle can use more other particles' useful information to strengthen the cooperation and competition between particles by the individual extreme value weighted mean.

II. PARTICLE SWARM OPTIMIZATION BASED ON THE GEESE

The standard PSO and various improved algorithms focus on how to make the particle swarm more effectively search the optimal solution in the solution space. But in the latter period of the search, particles tend to be identical, and this unification limits the search range of particles. To expand the search range, the number of particles must be increased in the particle swarm, or the particle's pursuit to the global optimum is weakened. Increasing the number of particles will lead to higher computational complexity of the algorithm. Reducing the particle's pursuit of the global optimum has the disadvantage that the algorithm is not easy to converge. The following improvements can be made to PSO [3]:

A. Improve PSO by using the flight characteristics of geese

In nature, the flight mode of geese is very efficient, and the flight distance of geese increases 72% more than solo goose. In flight of geese, leader goose flaps wings to produce vortex, and trailing companions can assist to fly. So the leader goose is the most laborious. The leader goose is the strongest goose and the other geese lines up in turn. Reference to the inspiration of the geese flight, the strength level of a goose can be regarded as the degree which particle is good or bad, namely historical optimal fitness value of particle. So all the particles will be sorted by the history optimal fitness value, select the best fitness value history optimal particle as the leader goose. The best fitness value of each particle is updated after each iteration, and then all particles are reordered.

The geese line up from front to back according to the historical optimum fitness value, each goose behind only follows its front the better goose flight. That is to say, the anterior goose's individual extreme is the global extreme of the behind goose ($p_{(i-1)d}$ replaces p_{gd}), and the global extreme of the leader goose is still its own individual extreme. This is the improvement of the global extreme value of PSO according to the flight characteristics of the geese. Speed formula is updated to:

$$v_{id}^{k+1} = \omega \times v_{id}^{k} + c_1 \times rand() \times (p_{id} - x_{id}^{k}) + c_2 \times rand() \times (p_{(i-1)d} - x_{id}^{k}) \qquad (1)$$

The advantages that the front optimal particle individual extreme replaces the global extreme value: All particles fly in more than one direction, avoid the tendency of particles to be identical, maintain the diversity of particles, and expand the search scope; but weakening the particles' chase to the global extreme lets algorithm not easy to converge.

In the geese flight, geese can push and cooperate with each other through the tail vortex, which is efficient because of group cooperation. The purpose of group cooperation is [4]: firstly, each individual can help other members of the group in the process of growing up; secondly, group cooperation can improve efficiency. In other words, each individual can provide information to the community, and each individual can assist other individuals in searching, just as multiple intelligences collaboration and competition. E.O.Wilson [5] argues that, at least in theory, in the process of mass search for food, each individual in the group can benefit from the new discovery of the group and the experience of all other individuals in the group. In the flight of geese, we think that the leader goose only relies on its own experience to make decisions. The behind geese not only rely on their own experience but also learn from other geese's experience, and its current value is a reference to the weight, the current fitness value represents the current state. So each goose individual extreme value except the leader goose is transformed to the weighted average value of individual extreme value and its present fitness value f(Xi).

$$P_a = \frac{\sum_{i=1}^{N} P_i \times f(X_i)}{\sum_{i=1}^{N} (X_i)} \qquad (2)$$

Improving p_{id} to p_{ad} has the following advantages: particles use more information to make their own decisions, which makes the algorithm to further reduce the probability of falling into local optimal; individual obtains more incentive, strengthen the cooperation and competition between particles, and accelerate the convergence speed.

Combination with the above two improvements, the speed and position formula of the new algorithm is updated as follows:

$$v_{id}^{k+1} = \omega \times v_{id}^{k} + c_1 \times rand() \times (p_{ad} - x_{id}^{k}) + c_2 \times rand() \times (p_{(i-1)d} - x_{id}^{k}) \qquad (3)$$

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \qquad (4)$$

The new algorithm refers the characteristics of the geese flight. On the one hand, the front optimal particle individual extreme replaces the global extreme value, so all particles fly in more than one direction, avoid the tendency of particles to be identical, and maintain the diversity of particles; On the other hand, the new algorithm makes each particle use more useful information of other particles, replaces the individual extreme with the weighted average value of individual extreme value and its present fitness value. Individual incentives become larger. The algorithm strengthens the cooperation and competition between particles. The combination of the two improvements balances the contradiction between the algorithm search speed and the algorithm accuracy.

B. Steps of GeesePSO

1) *Initialize the particle swarm:* give population size *M*, the solution space dimension *N*, randomly generate the location of each particle X_i, speed V_i.

2) *Calculate the current fitness value of each particle with the benchmark function f(X).*

3) *Update individual extreme:* evaluate the individual extreme value of each particle, compare the current value of the ith particle $f(X_i)$ with the fitness value of the particle individual extreme value P_i . If the former is excellent, update P_i, otherwise Pi is unchanged.

4) *Particle swarm sort:* all particles are sorted according to the historical optimal value (P_i fitness value), select the best history optimal fitness value particle as the leader goose, other geese turn back in turn.

5) *Calculate the new individual extreme:* the leader goose's individual extreme remains unchanged, calculate the new individual extreme (P_a) of other particles with the formula (2).

6) *Calculate the new global extreme:* the leader goose's global extreme remains unchanged, each goose behind takes the individual extreme of the front superior goose as its global extreme.

7) *Update speed and position:* update the velocity (V_i) and position (X_i) of each particle by formula (3) and (4).

8) *Check whether the stopping condition (maximum iteration algebra or minimum error threshold) is satisfied:* if it is satisfied, exit; otherwise, go to step (2).

III. Geometric Constraint Solving

From the point of view of artificial intelligence [6-7], the design problem is essentially a constraint satisfaction problem. Among the many design constraints, geometric constraint is the most basic. It is the basis for expressing other design constraints, and also a priority problem in constraint management and solution technology. The ultimate goal of solving a geometric constraint problem is to determine the specific coordinate position of each geometry in geometry. If the degree of edge generate (DEG) of a Geometry is less than its degree of freedom (DOF), the geometry can be determined by the location of the geometry in which it is bound.

In engineering applications, most mechanical designs come from sketches and existing graphics. In sketch design, the user initially does not care about the exact size of the graph, but roughly outlines the general shape of the part. The user may make minor improvements on the basis of the existing graphics. Size adjustment is very common, because size can determine the geometry of the parts. Size changes can produce different geometric shapes. The traditional interactive mapping method can give full play to the designer's ability. But after graphic production, it is difficult to adjust the size because it has not inheritance.

For the constraint problem, it can be formalized as (*E*，*C*) [8] (*E*= （e_1， e_2， ……， e_n）), it represents geometric elements, such as dots, lines, circles, etc.; *C*= （c_1， c_2， ... ， c_m） ， c_i represents the constraint between these geometric elements. Since a constraint corresponds to an algebraic equation, the constraint can be expressed as

$$\begin{cases} f_1(x_0, x_1, x_2, ..., x_n) = 0 \\ \quad ... \\ f_m(x_0, x_1, x_2, ..., x_n) = 0 \end{cases} \tag{5}$$

$X=（x_0，x_1，…，x_n）$

x_i is some parameters of the geometric element(e_i), for example the two dimensional point can be expressed as （x_1，x_2）. Constraint solving is to find the X formula (5).

$$F(X_j)=\sum_{1}^{m}\left|f_i\right| \tag{6}$$

If X_j satisfies $F(X_j)$=0, the X_j satisfies the formula (1). The constraint solving problem can be translated into an optimization problem, Only min($F(X_j)$)<ε is required, ε is a threshold. To improve the speed of the algorithm, we use the absolute value sum of the fi instead of the squares sum to represent the constraint equations. By formula (6) and using the GeesePSO to solve min($F(X_j)$)<ε (m=n is not required), the method can obviously solve the under- constraint and over-constrained problems.

IV. EXPERIMENTAL RESULTS

Figure 1 is the original design. Figure 2 is the new graphic that uses the GeesePSO after part of the size or angle are changed. From the diagrams, the user can modify the size value, and the system of cell membrane optimization algorithm updates graphics in real time according to the new size. That can easily create series parts and modify graphics. According to the above sketch, we compare the genetic algorithm PSO and GeesePSO.

TABLE I. COMPARISON OF THE EXPERIMENTAL RESULTS OF PSO AND GEEPSO

Algorithm	***Iterations***	***CPU occupancy time***	***Iteration Number of the best solution***
PSO	80	90	50
GeesePSO	40	50	40

It can be seen from table I. that the GeesePSO is used to solve the geometric constraint problem, and the algorithm can achieve better performance and better convergence than the other algorithms. The new algorithm not only has higher search speed, but also has higher convergence precision. It can balance the contradiction between the search speed and the accuracy.

V. CONCLUSION

Geometric constraint solving is the core of parametric design. The quality of geometric constraint solving is the key to the parametric design system. In this paper, the constraint equations of geometric constraint problems are transformed into optimization models, the problem of constraint solving is translated into optimization problems. As one of the representative methods of swarm intelligence, particle swarm optimization algorithm provides a new solution for nonlinear, non-differentiable and multi-peak complex optimization problems, but it is easy to fall into local optimum and divergence. In this paper, an improved PSO algorithm is proposed by referring to the flight characteristics of the geese.

On the one hand, the global extreme value transforms to the individual extreme of the front optimal particle, all particles fly in more than one direction. That avoids the tendency of particles to be identical, and maintains the diversity of particles. On the other hand, the new algorithm makes each particle use more useful information of other particles, replaces the individual extreme with the weighted average value of individual extreme value and its present fitness value. Individual incentives become larger. The algorithm strengthens the cooperation and competition between particles. The combination of the two improvements balances the contradiction between the algorithm search speed and the algorithm accuracy. Experimental results show that GeesePSO has higher convergence accuracy, faster convergence speed, better global search capability, and a proper balance between detection and development capabilities in geometric constraint solving.

REFERENCES

[1] Yuan Bo. The research and implement of geometric constraint solving［D］. Beijing:Tsinghua University，1999.

[2] Holland J H. Adaptation in natural and artificial systems［M］ Cambridge: MIT Press，1975. .

[3] Liu Jin-yang，Guo Mao-zu，Deng Chao. GeesePSO: an efficient Improvement to particle swarm optimization ［J］. Computer Sci-ence，2006，33(11):166-168.

[4] Beekman M, Rantnieks FLW. Long-range foraging by the Honey-bee, Apis Mellifera L. Functional Ecologicy, 2000, (14): 490-496

[5] Wilson E O. Sociobiology: The New Synthesis [M]. Cambridge: Belknap Press, 1975.

[6] Shi Zhi-liang，Chen Li-ping. A simplified iterative algorithm to solve geometric constraints[J]，Journal of Computer-Aided Design & Computer Graphics，2006，18(6):787-792. .

[7] Sun Wei，Ma Tie-qiang，Huang Yu-jun. Research on method of constraint conversion in feature-based data exchange between heter-ogeneous CAD systems[J] . Journal of Mechanical Science and Technology，2009，23(1):246-253. .

[8] Liu Sheng-li，Tang Min，Dong Jin-xiang. Geometric constraint satisfaction using genetic simulated annealing algorithm [J]. Journal of Computer Aided-design & Computer Graphics，2003，15 (8):1011-1029.

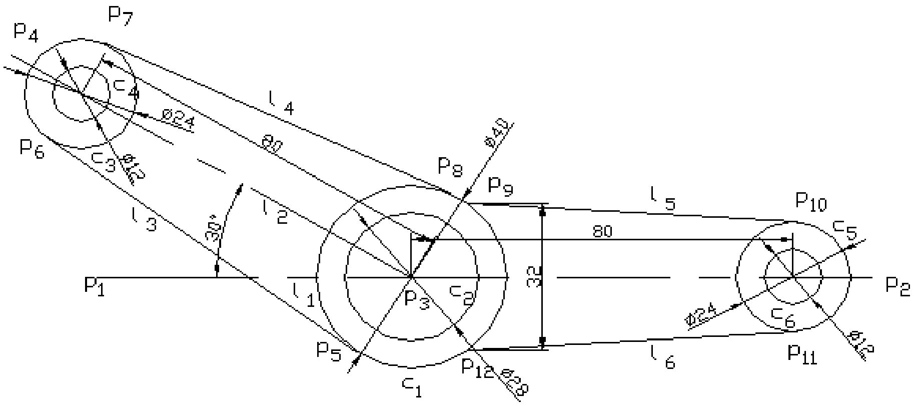


Figure 1. Original design

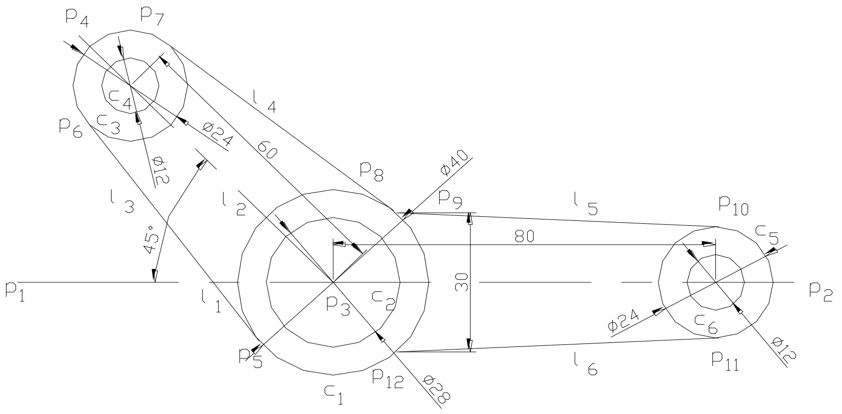


Figure 2. New graphics designed by GeesePSO algorithm

The Application of Human-Computer Interaction Idea in Computer Aided Industrial Design

Zhang Liang
Mechanical and Electrical Engineering Institute
Qingdao Binhai University
Qingdao, China
e-mail: 76201691@qq.com

Zheng Li-nan
Mechanical and Electrical Engineering Institute
Qingdao Binhai University
Qingdao, China
e-mail: 1021090387@qq.com

Zhao Jian
Mechanical and Electrical Engineering Institute
Qingdao Binhai University
Qingdao, China
e-mail: 84310626@qq.com

Li Nan
Mechanical and Electrical Engineering Institute
Qingdao Binhai University
Qingdao, China
e-mail: 691525115@qq.com

***Abstract*—The interactive, experiential, real-time, efficient and comprehensive features of human-computer interactive technology in virtual display make it possible to virtualize products with panoramic, instantaneous and experiential features. The computer aided design based on human-computer interactive technology, through the design and present from shape, color and structure of products, enables designers to pre-design in advance to avoid the mode of production then design in before, which can be a good reduction in industrial production and manufacturing costs, at the same time, it is possible to evaluate the possible problems in the design of industrial products and some defects in products, so as to improve to achieve a good design process.**

Keywords-Human-computer interaction; Virtual display; Experience; Computer aided design; Products design

I. INTRODUCTION

With the rapid development of digital information technology, the research on the theory and method of modern industrial design has made great progress. The technology of computer aided industrial design fully involved in the field of industrial design and the idea of human-computer interaction makes people have an urgent demand for a dynamic, three-dimensional, real-time and realistic display experience. So, the computer aided design based on the idea of human-computer interaction has injected new vitality into industrial design.

The unique interactive features of computer technology gives the product display a strong visualization. Designers can quickly demonstrate the scheme designed at the conceptual stage by computer and make of three-dimensional model, so that the designers can real-time visualize the various parts of the works in the design process, find out the shortcomings to modify in the next step, until to see the results of design intuitively. The whole process can be seen through virtual display platform, so that customers can see in advance the real effect of the design after the work is really finished and then make suggestions for modification of the design proposal. The combination of human-computer interaction idea and computer-aided design shows for people the design process ,the actual image and application effect after the completion of the design comprehensively, specifically and in real time.

II. THE CHARACTERISTICS OF HUMAN-COMPUTER INTERACTION IN COMPUTER AIDED DESIGN ENVIRONMENT

A. Interactivity

Interactivity refers to the degree of operability and responsive to user actions of the display in a virtual environment. In the virtual product exhibition design based on human-computer interactive technology, the computer can generate a virtual environment consistent with the characteristics of the product and the original design, so the users can take the initiative to receive the information conveyed by both the virtual environment and the digital display [1].Virtual display design is not a static closed world, but an interactive and open system, which can affect the users or be affected through the design, control, management and device mobilization.

B. Experience

The display of design based on human-computer interactive technology can create a real virtual environment for viewers, and through the three-dimensional digital model by computer, so that people can receive a comprehensive experience in visual, auditory, tactile and other sensory organs.

C. Immediacy and efficiency

The display of design based on human-computer interactive technology can directly across the time cost consumed by the traditional display, directly display the latest products and their characteristics to the users through the digital technology, multimedia tools and virtual reality network platform. The target customer groups and users of the product can receive and browse the information in a timely manner, which reflects the efficiency of virtual display.

D. Comprehensiveness

The virtual display design system based on user needs can build a product model with 3D entity to fully and accurately show the structure and performance of products, at the same time, transmit the information to the user in a timely manner through the network, so that users can fully grasp the product information in the virtual display system. In general, the product information mainly includes the product's 3D model information and performance information. The former includes static information, such as appearance, color, structure, material and so on, the latter refers to the dynamic operation of the product information, such as the use of the characteristics of the process and the performance and status. The most important thing is that these physical information will be transformed into digital information in the virtual display design, so that the users can grasp the product information fully and accurately.

III. THE OVERALL FRAMEWORK OF COMPUTER AIDED INDUSTRIAL DESIGN SYSTEM

The development trend of computer aided industrial design technology mainly focuses on product design research, computer application technology research, and on the whole, provide technical support for the whole process of industrial design. The overall framework of the computer-aided industrial design is shown in Fig. 1.

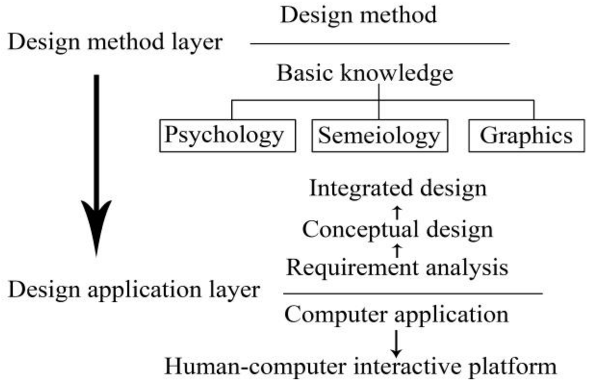


Figure 1. Computer aided industrial design overall framework

The design methodology layer mainly includes two layers: the knowledge base and the design method. The design method layer mainly refers to how to carry out the activities of industrial design on the basis of computer aided industrial design, and to seek and improve the guiding theory and method of industrial design [2].The continuous exploration and innovation of the knowledge base layer is beneficial to human-computer interaction, user requirements, artificial intelligence, and more scientific modeling design methods.

The design application layer is embodiment of the entire computer-based industrial design process. Computer aided technology for technical support of industrial design is mainly reflected in the demand and industrial design development trend of users [3].Among them, the user needs design, mainly to solve how to integrate the user into the loop design, and also consider the user needs of the specific and targeted.

The problem that urgent needs to be solved in designing application layer is how to design interface for user's requirement design, detailed design and conceptual design.

IV. PRODUCT DESIGN PROCESS UNDER THE IDEA OF HUMAN-COMPUTER INTERACTION

The globalization of computer information network has occupied an unshakable position in people's life and work, and the modern society has changed from "technology as the foundation" to "information technology as the foundation" [4], which indicates that the society is advancing with the times, but also brings some problems to industrial designers.

On the premise of human-computer interaction, all the connections between human and computer need to be maintained by human-computer interface. Human computer interface is the link between human and computer interaction and connection. It is based on the information tradition of dealing with "people and things". It not only studies the use of computers, but also emphasizes the importance of human beings [5].The purpose of man-machine interface is that it deals with the relationship between industrial designers and computer software and hardware, studies the design of man-machine interface model, the design of virtual interface, the design of multi-user and multi-sensory interface, and so on, so asto provide a good technical foundation for industrial design.

In order to effectively use the computer to assist the design, and the design information accurately conveyed to the customer, when the designers deal with every detail of the product in the use of computers as a means ,they must ensure that customers and industrial products be accessible communication between the interaction .Therefore, human-computer interaction is particularly important in the process of industrial design. At present, in the computer-aided industrial design, human-computer interaction is mainly reflected in the product design process based on conceptual design, personnel collaboration, product design, virtual display, design feedback and product improvement.

The human-computer interaction in the product design process is shown in table 1.

TABLE I. HUMAN-COMPUTER INTERACTION IN PRODUCT DESIGN PROCESS

The process	**Human - computer interaction target**		
	computer	***counterparts***	***customer***
conceptual design	YES		
personnel collaboration	YES	YES	
product design	YES		
virtual display	YES	YES	YES
design feedback	YES	YES	YES
product improvement	YES		

A. Conceptual Design

The conceptual design of industrial products is to explore the combination of the functions and structures of industrial products, so as to get the best combination of the product in the design stage, which is a design process that designer give full play to the imagination and creativity. Computer aided industrial design technology will provide technical support for conceptual design of designers throughout the process, that is, from the design inspiration to the specific design implementation will provide aided design help.

Although computer aided industrial technology allows the concept of designers to be represented by two-dimensional or three-dimensional graphics, but the latter part of the modification and maintenance is not very convenient, therefore, at this stage, human-computer interaction based sketch aided design system can help designers get rid of repeated manual drawing work, and can easily express innovative design [6].

B. Personnel Collaboration

Modern industrial design has long been out of the traditional manual design, but the design structure, aesthetic point of view, interactive information and design resources put forward a higher requirements. In the specific process of industrial design, designers of different styles and different academic backgrounds should be gathered together to cooperate with each other to provide technical support for industrial design. In the concept of human-computer interaction, computer aided industrial design technology can help designers in different places jointly design and develop industrial model by using synchronous and asynchronous technology, so that designers can easily share industrial design information and design inspiration, thereby significantly improving the efficiency of industrial design.

C. Product Design

In the process of product design, virtual assembly and virtual simulation can effectively reduce or even avoid errors. The maintainability, configurability and compatibility of industrial products are key to the constant error of the designers. In the past, during the final assembly of the product, it was not until a long period of time that to find the parts in the assembly were broken or even scrapped. Virtual assembly refers to the industrial product design stage that designers use the way of human-computer interaction directly three-dimensional assembly for industrial products, and in the assembly process make designer's creativity to maximum. At the same time, in this three-dimensional three-dimensional effect, designers observe the industrial products as in real life which can be the fastest speed to accurately find the flaws in technological design process.

The application of virtual simulation technology in industrial design refers to the accurate judgment and design of human-computer interaction through the technology of virtual simulation in computer system [7].Through the virtual simulation technology can achieve a goal that operate and test each of the design process to help the designers faster and more accurate to complete or modify the design scheme, reduce unnecessary time and mental labor, strengthen the technical communication with the design members, explore the resources, and improve the design speed of industrial products.

Today's industrial products not only meet the user's normal functional requirements, but also meet the aesthetic needs of users, which requires designers to increase the importance of the product's appearance, material, overall structure and color matching, and even can use the computer-aided industrial design technology for all aspects of the above to achieve the ideal product design by analogy, analysis, selection and arrangement [8].

D. Virtual Display

Design and display based on the idea of human-computer interaction is designed to meet the needs of users and improve the quality of interaction. The interactive quality is decided by two aspects: One is usability, the other is the user's experience. This paper analyzes the important role of human-computer interaction idea in industrial design through a car showcase [9].

Virtual display based on human-computer interaction idea does not appeal to users only through pictures, text or video music, but interactive design plays a decisive role in it. In order to make the experimenter fully interact with this presentation, the interaction of the whole framework includes perspective interaction, color material interaction, performance interaction and feedback, driving mode interaction and so on. Its structure is shown in Fig. 2.

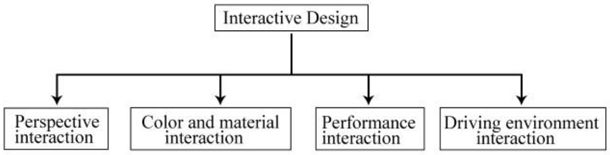


Figure 2. Interactive system for vehicle virtual display design

Take the perspective interaction as an example. In order to give the user the most real and natural visual experience, there is a variety of perspectives for users to choose, including control the angle of view through draging the mouse and local features [10].People can also add more navigation on this basis, such as through the keyboard to control the rotation of the field of view, and create an external interface, through relative devices to capture the experience of the head or body positioning, according to the experience of the real-time perspective angle replacement, together with three-dimensional glasses and other external devices, restore the user a real and easy to use man-machine interface [11].

In addition, different configurations will be brought into the driving mode, so that users can really feel the appearance, interior and performance of the car when driving, which will help users gain a sense of experience.

The virtual environment of the car is divided into browsing the environment and driving environment, browsing is divided into multi-angle view and the car driving angle; driving environment for the third-person camera perspective, the perspective can clearly see the car driving state, including wheel steering and body with the ground ups and downs.

In the browsing environment, user control of the camera as the free navigation mouse interaction: press the left mouse button and move the mouse to control the angle of rotation, if fast rotation also joined the buffer visual effect, that is, the mouse will not immediately stop after the release of rotation, it will be in the original rotation on the basis of a inertia, and gradually stop.

The activation key of the driving mode is the E key on the keyboard, which is located in an open mountain environment where scenes and cars are given a very real physical property, including gravity, collision, and so on. In this mode, the user can visually observe the state of the car, fully feel the flexibility of the car. At the same time, the given physical attributes can be adjusted according to the actual conditions of the car, so the final driving state of the car is more realistic.

E. Design Feedback

In virtual display, in order to facilitate the communication between the designer and the end user, can provide special modules for users to choose the configuration freedom and feedback. In this case, wheels and car interiors are taken as examples.

The user switches through the selection button of the main interface, in addition to real-time rendering in appearance, it will also change the background database in real time and feedback the data to the designer. For example, in the color and material interaction module, parts related to color and material change including body, ceiling, seat and so on. These interactive parts are triggered by buttons.

When the customer clicks the button, the computer system retrieves the current color information from the background database and passes it to the main interface's information statistics panel to display the color information visually. At the same time, this information can be saved as a final customer selection scheme and submitted to the designer to achieve personalized customization. This process is shown in Fig. 3.

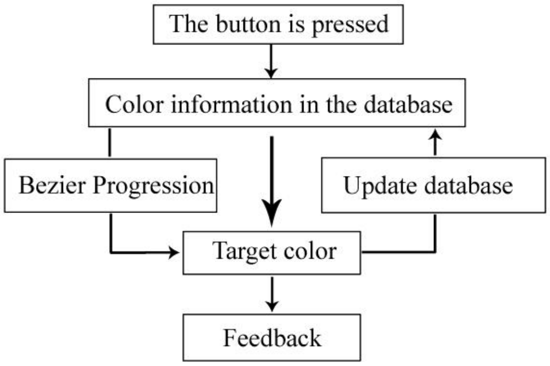


Figure 3. Sketch map of color and material interaction program

F. Product Improvement

In the future, the resonance of industrial products and consumer sentiment will become one of the important factors for consumers to choose industrial products, simply paying attention to the aesthetics of industrial products has not been suitable for the development of computer aided design [12].The application of computer-aided design in product design should be based on the principles of ergonomics, human nature design and sustainable design, in the process of computer aided design, more humanized design elements are added, and the interaction design and emotional design of products are taken as the main factors. Therefore, the designer should follow the customer feedback and computer system information records, re-examine the product modeling, color, materials, structure and other elements, take human-computer interaction and experience design as the guide to improve and optimize the products.

V. CONCLUSIONS

From the industrial design itself, with the continuous development of artificial intelligence, virtual simulation and other technologies, designers' thinking will also undergo major changes. The CAID of human-computer interaction mode will be the inevitable trend of the future industrial design, and the more humane, fast and real human-computer interaction will be the inevitable result of human-computer interaction in CAID.

In the field of computer technology and human-computer interaction, more bold and innovative concepts have been put forward. Language identification, infrared remote sensing, video capture and so on, will inevitably bring more technical improvements to human-computer interactive technology. The new human-computer interactive technology has been emerging, and it can bring more scientific and technological innovations and expectations. In the future computer industry design, human-computer interaction should strengthen the "human centered", "harmony" "unified" interactive model, so as to improve the efficiency of human-computer interaction.

REFERENCES

[1] LIU Qi-wen,qiu feng.Research on display in industrial design based on interactive computer technology[J].Journal of Wuhan University of Technology,2008,30(9):163-164.

[2] ZHANG Tian-cheng.The design and research of computer aided industrial design system[J].Journal of Liao Ning Institute of Science and Technology,2016,18(4):7-9.

[3] ZHOU Su,WANG Wen.Human-computer interactive technology[M].Tsinghua University Press,2016.06.

[4] DU Gang-ji.Elementary introduction to industrial design and application of computer technology[J].Industrial Design,2015(04):111-112.

[5] HUI Qi-xun Computer-aided industrial design of human interaction[J].Computer Knowledge and Technology,2012,08(17) 4244-4245.

[6] LI Ying.Exploration of human-computer interaction based on computer technology[J].Information and Computers (Theoretical Edition),2016(7):59-60.

[7] HUANG Xian-qiang.The applied research of interaction design in industrial design[D].Qilu University of Technology,2014.

[8] ZHANG Jian-hui,TAN Run-hua,ZHANG Zheng-yan,etal.Research on the integration of product concept design and particular design driven by CAI technology[J].Journal of Mechanical Engineering,2016,52(5):47-57.

[9] TAN Hao,ZHAO Jiang-hong,WANG Wei.Vehicle human machine interface design research[J].Journal of Automotive Engineering,2012,02(5):317-318.

[10] TENG Jia-ni.Analysis of core principles of interactive design based on user experience[J].Art and Technology,2013,(10):289-289.

[11] SUN Xiao-hua,FENG Ze-xi.Interaction design for wearable devices[J]. Decoration.2014,(2):29-31.

[12] ZHU Wei.The development situation and trend of computer aided industrial design[J].Journal of Hubei University for Nationalities(Natural Science Edition),2013,31(2):222-224.

Research on Low Voltage Power Line Carrier Communication Simulation Software

Ye Jun
Chong Qing Electric Power Research Institute, Chong Qing, China
gtovictor@163.com

Sun Hongliang
Chong Qing Electric Power Research Institute, Chong Qing, China
cqepshl@sina.com

Li songnong
Chong Qing Electric Power Research Institute, Chong Qing, China
lxpecolicee@163.com

Hou Xingzhe
Chong Qing Electric Power Research Institute, Chong Qing, China
cqhhxz@163.com

***Abstract*—The use of semi-physical simulation platform for school laboratory and classroom building scene a lot of field measurements, analysis of power line channel transmission characteristics of the different environments, and features a three-dimensional graphic display of variable power line channel under different scenarios. At the same time in order to improve the efficiency of the data analysis, the use of MATLAB simulation software design based on off-line data analysis software GUI interface. It receiving end signal processing module integrated into the GUI interface, to implement graphical data analysis and processing, and at the same time from multiple dimensions of the channel parameters display.**

Keywords-Adaptive impedance matching; Bit Error rate; Communication equipment; Continuous adaptation

I. INTRODUCTION

With the advent of the information age, low voltage power line carrier communication technology because of its communication lines without additional laying, and widely distributed, so it has broad application prospects in the room automation. However, as a low-voltage power line is mainly used for power transmission lines of the communication environment is very complex, a lot of power line carrier communication technology application development is still in the experimental stage, there are still many difficulties to be resolved.

Low-voltage power line load work status changes, resulting in time-varying characteristics of strong channel transmission characteristics complex, concentrated reaction in the input impedance, signal attenuation and noise three areas. Specific performance input impedance changes, the communication device without having to constantly and line impedance effectively matched, resulting in the carrier signal energy can not achieve the desired output; the same time as the power line complex network structure distribution, so severely attenuated phenomenon signal transmission process occurs, resulting in the signal waveform severe distortion, certain difficulties to the reception carrier signal; and the channel there is a lot of background noise, seriously affect the accuracy of the information-data transfer, resulting in serious errors received data. Impedance characteristic low-voltage power line noise attenuation characteristics and also with the dramatic changes occur in different times and places, has a strong randomness. As in the above three properties in the evening peak period and late trough measured at the same location will be out

Now a huge difference, as the electricity network at the same time to test different cells also showed very different. Because of the complex nature of these low-voltage power line channel, many R & D personnel carrier communication device discovery has been developed in the field of communications equipment running very unstable. Experimental data at different times or in different locations will have a significant difference, and thus to the development and testing of experimental evaluation of communications products brought many difficulties. For this reason, the concept presented here to build a low-voltage power line carrier communication comprehensive experimental test simulation platform, attempts to provide a unified common test environment for R & D personnel and low voltage power line carrier communication, the communication product development as much as possible in a laboratory experiment into, rely solely on the scene to overcome the situation. Establish such a system, but also conducive to communication equipment factory inspection, evaluation of performance, and improved means of communication. At the same time, given the current low-voltage power line channel characteristics analysis and modeling at home and abroad mostly in theoretical research level, so here uses software and hardware combination of methods to establish the analog low-voltage power line carrier communication channel characteristics of the test simulation platform for power line carrier communication technology the research work, but also helpful.

II. CHANNEL CHARACTERISTICS OF LOW VOLTAGE POWER LINE CARRIER COMMUNICATION ANALYSIS

Designed for low-voltage power lines are used to transmit power frequency 50Hz power, does not need to transmit high-frequency communication signals and perform special consideration. So it's topology and physical characteristics there is a huge difference between traditional communication transmission media (twisted pair, coaxial cable, fiber, etc.), resulting in low voltage power line communication channel environment is very bad, the channel characteristics quite complex. Mainly for line input impedance changes, it is difficult to match with the transceiver; high frequency signal attenuation serious and difficult to detect output signal; a large number of low-voltage power grid noise sources exist, communication environment interference large. Through the study of low-voltage power supply system power line channel characteristics, we find the basic characteristics of the channel are as follows [12 - 14]:

(1) The power supply system is in the form of parallel distribution to each user, the system is open, vulnerable to interference from user factors.

(2) power system frequency of 50Hz, all the power supply wires are made of low-frequency, low-cost aluminum or copper production, coupled with the ease of network structure and so will increase the power supply system of distributed capacitance, frequency variation .

(3) power supply system will vary with the number of users and electrical access and random changes occur.

(4) interference system load and the environment by large, access and disconnect the inductive load, thyristor surge devices, switching power supply electrical noise, power failure caused by tripping and fuse so the system have a greater interference.

(5) Since the random network structure of the power supply system, and power supply system and the load characteristics of the line itself random access, resulting in a mismatch of the characteristic impedance. To sum up, the channel transmission characteristics of the power line complex concentrate reaction in the input impedance, signal attenuation and noise three areas. The following detailed analysis will form the original low-voltage power line channel basic characteristics, the input impedance characteristics, transmission attenuation characteristics and channel noise characteristics demonstrated by the changes in the characteristics.

$$y_B^{plc} = h_{AB}^{plc} x_1 + h_{DB}^{plc} x_2 + n_B^{plc} \tag{1}$$

$$y_C^{plc} = h_{AC}^{plc} x_1 + h_{DC}^{plc} x_2 + n_C^{plc} \tag{2}$$

$$\begin{aligned} y_C^{wl} &= h_{BC}^{wl} y_B^{plc} + n_C^{wl} \\ &= h_{BC}^{wl} h_{AB}^{plc} x_1 + h_{BC}^{wl} h_{DB}^{plc} x_2 + h_{BC}^{wl} n_B^{plc} + n_C^{wl} \end{aligned} \tag{3}$$

$$I_{SDF} = \begin{cases} \frac{1}{2}\log_2(1+2\gamma h_0), & h_2 < \gamma_{th} \\ \frac{1}{2}\log_2\left(1+\gamma h_0+\gamma h_2\right), & h_2 \ge \gamma_{th} \end{cases} \tag{4}$$

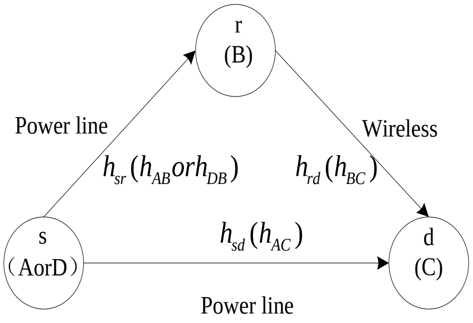


Figure 1. Basic structure of proposed PLC system

III. POWER LINE TRANSMISSION MODEL BASED ON MULTI-PORT

Features summarized as follows:

(1) channel attenuation is mainly due to the coupling attenuation and line attenuation caused in two ways. Coupling attenuation by Output impedance of the signal transmission circuit and the power line input impedance mismatch caused. Attenuation refers to the transmission line Loss of energy input signal lines on the power line, its causes are many, including power lines Network complexity, a number of the access node, and the channel there are many nodes in the impedance mismatches and the like. Low-voltage power line carrier communication channel tend to have the characteristics of the multipath channel, will inevitably lead to more signals Path propagation, resulting in attenuation.

(2) As the frequency increases, the attenuation of the signal will also increase. However, in some special frequency bands, by On the impact of reflection, resonance and transmission line effects such as multipath, attenuation will fluctuate at a specific frequency Change.

(3) The degree of signal attenuation generally speaking, is proportional to the distance of signal transmission. However, due to the power line carrier communication simulation software on the line parameters of the power line, load parameters, noise parameters, and add in the signal parameters power line set by software algorithm simulation to achieve a low-voltage power line communication channel input impedance characteristics, transmission attenuation and interference noise and other characteristics of the power line itself changes, to provide a power line carrier communication with the actual situation close to the channel environment.

Facilitate carrier communication personnel in the preliminary design stage through simulation communication simulation software for analysis to determine the transmission effect of the overall design, but by observing the distortion of the situation to send the signal waveform and the received waveform, to study the line parameters of the power line, load, signal frequency, impact of noise on the power line carrier communication, and try to influence the different communication modulation and demodulation algorithm for data transmission errors, thus providing an operation on a PC, the software platform and simulation environment for power line carrier communications test experiments.

Overall design of power line carrier communication simulation software can be represented by a data flow diagram shown in Fig.3-1. Firstly set the parameters of power line carrier communication by the user, which includes the channel line parameters, namely the length of distribution lines, cable type transmission line resistance, line capacitance, line inductance other items, including signal parameters, noise signal parameters and load setting parameters. Signal parameter setting signal includes a type, amplitude, phase, frequency, the noise signal includes noise intensity parameter settings, the type of coupling point location, the load parameter set includes mainly the form and size of the load.

Choose a variety of topologies channel is also provided by the user parameters to achieve, so the user must select the parameter values before the simulation according to their own circumstances. After completion of the parameter setting parameter data into a unified data store to call during the simulation calculations. When Parameter settings can also be saved in a file, data will be sent to save the parameter settings of the module to save for the next simulation using the same set of parameters, you can call the saved file directly, to avoid duplication of effort.

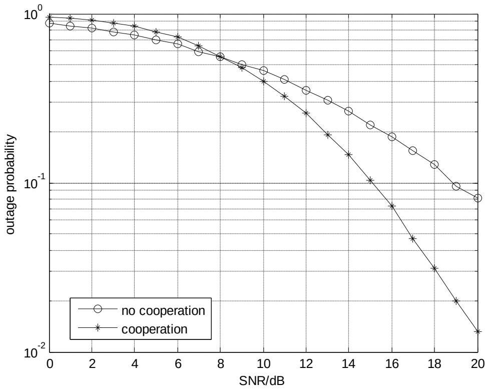


Figure 2. utage probability of cooperation system and no cooperation system with the SDF cooperation model

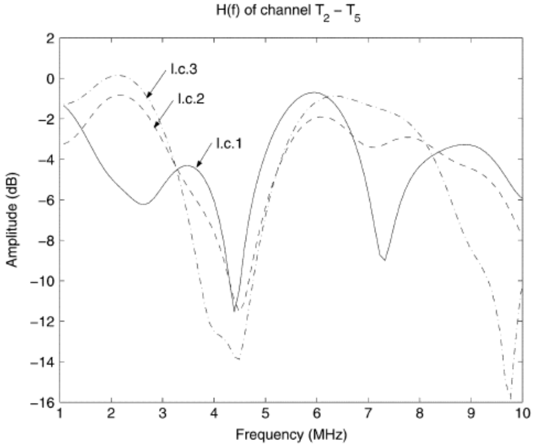


Figure 3. frequency response of the channel in the 1-10MHz band

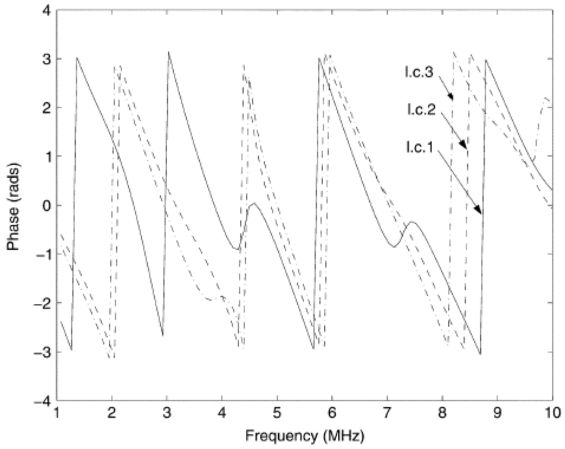


Figure 4. RMS delay spread for channel T2-T5

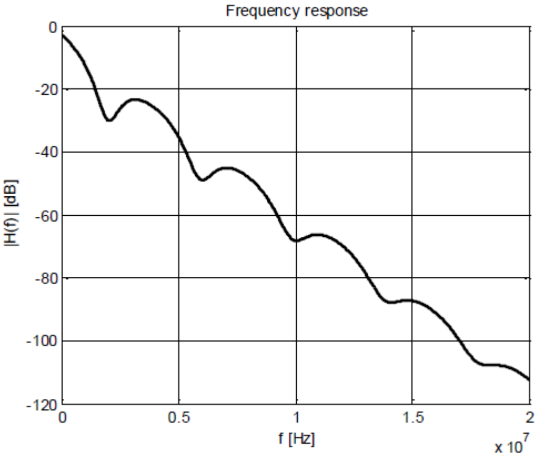


Figure 5. simulation of the sample network

IV. CONCLUSION

Setup procedure under the data flow good parameter level is as follows: First, the user set up the channel model of the channel characteristics of the test analysis, including input impedance characteristics, the transmission attenuation and channel noise characteristics, so that the test results observable meets expectations channel performance requirements, if greater access should return to the previous channel reset line and load parameters, until satisfied. Then set the data good noise parameters and signal parameters added to the channel model, communication simulation. Finally, on the one hand by the output waveform module can display waveform signal after transmission through the channel distortion situation; on the other hand can be observed through the error rate of the entire power line carrier communication after the data transfer process, to verify the feasibility of communication algorithms.

ACKNOWLEDGMENT

This work is supported by State Grid Corporation of China technology project and cstc2016jcyjA0214

REFERENCES

[1] P. Langfeld and K. Dostert, “OFDM system synchronisation for powerline communications,” in Proc. 4th Int. Symp. on Powerline Communications and its Applications, Limerick, Ireland, 2000, pp. 15–22.

[2] M. Busser, T. Waldeck, and K. Dostert, “Telecommunication applications over the low voltage power distribution grid,” in Proc. IEEE 5th Int. Symp. Spread Spectrum Techniques & Applications, vol. 1/3, Sun City, South Africa, 1998, pp. 73–77.

[3] O. Hooijen, “A channel model for the residential power circuit used as a digital communications medium,” IEEE Trans. Electromagn. Compat., vol. 40, pp. 331–336, 1998.

[4] G. Threin, “Datenübertragung über Niederspannungsnetze mit Bandspreizverfahren, Fortschrittberichte VDI, Reihe 10,” VDI-Verlag, Düsseldorf, 156, 1991.

[5] J. Barnes, “A physical multi-path model for power distribution network propagation,” in Proc. 1998 Int. Symp. Powerline Communications and its Applications, Tokyo, Japan, Mar. 1998, pp. 76–89.

[6] A. Dalby, “Signal transmission on powerlines—Analysis of powerline circuits,” in Proc. 1997 Int. Symp. Powerline Communications and its Applications, Essen, Germany, Apr. 1998, pp. 37–44.

[7] M. Karl, “Möglichkeiten der Nachrichtenübertragung über elektrische Energieverteilnetze auf der Grundlage Europäischer Normen, Fortschrittsberichte VDI, Reihe 10,” VDI-Verlag, Düsseldorf, 500,1997.

[8] M. Zimmermann and K. Dostert, “A multi-path signal propagation model for the powerline channel in the high frequency range,” in Proc. 3rd Int. Symp. Powerline Communications and its Applications, Lancaster, U.K., 1999, pp. 45–51.

[9] H. Philipps, “Modeling of powerline communication channels,” in Proc. 3rd Int. Symp. Powerline Communications and its Applications, Lancaster, U.K., 1999, pp. 14–21.

[10] HRASNICA, H., HAIDINE, A., LEHNERT, R. Broadband Powerline Communications Network Design. [s.l.] : Willey , c2004. 275 s. ISBN 0-470-85741-2

[11] Babic, M.; Hagenau, M.; Dostert, K.; Bausch, J. Theoretical postulation of PLC channel model. Open PLC European Research Alliance (OPERA).2005

Visualization Analysis of NoSQL Research Field Based on SCI by CiteSpace Ⅴ

Ming He, Ying Zhang
School of Management
Beijing Normal University, Zhuhai
Zhuhai Guangdong, China
E-mail: heming2018@foxmail.com

Jianning Zhang
School of Management
Beijing Normal University, Zhuhai
Zhuhai Guangdong, China
E-mail: zhang1108545@gmail.com

Pixian Zhao
School of Management
Beijing Normal University, Zhuhai
Zhuhai Guangdong, China
E-mail: zhaopixian@bnuz.edu.cn

Yingxin She
School of Management
Beijing Normal University, Zhuhai
Zhuhai Guangdong, China
E-mail: syxingin@163.com

Yongjun Wu
School of Management
Beijing Normal University, Zhuhai
Zhuhai Guangdong, China
E-mail: jackripperwu@foxmail.com

Qike Jiang
School of Management
Beijing Normal University, Zhuhai
Zhuhai Guangdong, China
E-mail: jiangqike0504@foxmail.com

***Abstract*—NoSQL is one of the technical trends that rises in this context in the Web 2.0 Era. With the aim to explore the research status and development trends related to NoSQL technology, articles between 1998 and 2016 were collected from Thomson ISI's SCI. After the analysis by using CiteSpace Ⅴ, the pivotal documents related to NoSQL, as well as institutions, co-citation patterns, research hotspots and frontiers, etc., were visualized and identified.**

Keywords-NoSQL; Visual analysis; CiteSpace; SCI-E; Database

I. INTRODUCTION

With the continuous development of Internet technology, vast amounts of data have emerged in all areas of social life and scientific research. Faced with such a large-scale data, particularly in the SNS type of high concurrency scenarios, it has been a bit of powerless to store and query the users' dynamic data by using relational database [1]. There are a lot of advanced data management technology to alleviate this problem, NoSQL (originally referring to "non SQL", "non relational" or "not only SQL") is just one of the technical trends that rise in this context.

The databases like "NoSQL" have existed since the late 1960s, but did not obtain the "NoSQL" moniker until a surge of popularity in the early twenty-first century [2], triggered by the needs of Web 2.0 companies such as Facebook, Google, and Amazon.com [3, 4].

Johan Oskarsson reintroduced the term NoSQL in 2009 when he organized an event to discuss "open source distributed, non-relational databases". The name attempted to label the emergence of an increasing number of non-relational, distributed data stores, including open source clones of Google's BigTable/MapReduce and Amazon's Dynamo.

The main objectives of this study are to identify the current development status, trends and frontiers, to find core researchers as well as their co-citation situation, and to detect the pivotal documents in the research area of NoSQL from 1998 to 2016.

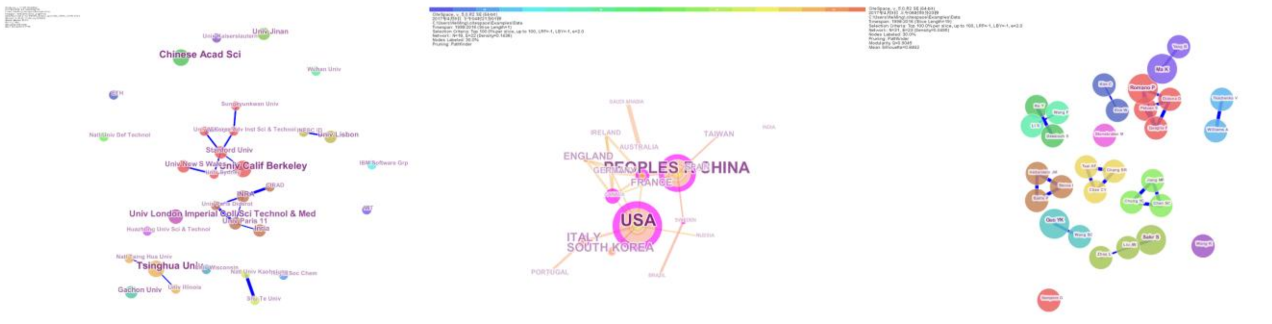
II. DATA SOURCES AND RESEARCH METHODS

A. *Data Source*

The data analyzed by this study is from Thomson ISI's SCI (Web of science in the Science Citation Index Expanded Edition). The time of collecting data is March 31, 2017. The author set the search mode to advanced search with the following formula: "TS = ((nosql) OR (non-relational database $) OR (nonrelational database $))". The timespan is set to 1998-2016 and the language is English. A total of 144 records include authors, titles, keywords, abstracts, and cited references.

B. *Research Tools*

In 2004, CiteSpace was first developed by Chaomei Chen to facilitate the visual analysis of trends, patterns, and critical changes in a changeable information environment [5]. In CiteSpace, Timeline views and time-zone views display the publication time and peak time of articles and terms, Cluster views is node and link diagrams, where the nodes present author, institution,



The institution network related to NoSQL researches (1998-2016)
(a)

The country network related to NoSQL researches (1998-2016)
(b)

The author network related to NoSQL researches (1998-2016)
(c)

Figure 1 The network related to NoSQL researches (1998-2016)

Country, term, keyword, cited reference, cited journal, and so on [6, 7]. The node size represents the overall citation frequency. Link represents co-citation or co-occurrence, the line's thickness represents the strength proportion of co-citation or co-occurrence. Each color corresponds to a time slice following the legend bar above the visualization area. However, if a node has a purple ring, which means that the node has a high betweenness centrality and tends to be strategically important in terms of the macroscopic structure of a new work. Those nods with high betweenness centrality are called pivotal points or turning points; if a node has a red ring; it means the node has burst in one of its attributes, notably citations [8].

Therefore, researchers can easily analyze the trends, patterns, and critical changes by studying the size, color of nodes, and links of colorful network. The version of the CiteSpace 5.0. R2 SE was the main research tool used in this paper.

III. DISTRIBUTION OF NOSQL RESEARCH FIELDS ANALYSIS

Distribution on research field analyses, including institution co-occurrence, country co-occurrence, author co-occurrence, are used to reveal the development status of NoSQL from different dimensions.

A. *Institution Co-occurrence Network Analysis*

Node type on the interface of CiteSpace was selected as the network node for the analysis of institutions. Because the total of institutions was fewer, time slicing was set to 19, which is the maximum value. Through running CiteSpace, then we can get holding the Fig. 1 (a) with 32 institutions and 17 links. Each country issued a relatively average number of articles. Among these institutions, Chinese Acad Sci, Tsinghua Univ, and Univ Calif Berkeley issued the largest number of documents.

B. *Country Co-occurrence Network Analysis*

The author set the node type to Country and time slicing to 1, and then run CiteSpace. A network consisting of nodes represented collaborating countries is presented in Fig. 1 (b).

In the NoSQL field, the United States has the greatest advantage, living in the world's first, and cooperation with other countries or regions more closely. China ranked second, significantly more than other countries and regions, but relatively less in terms of cooperation, followed by Canada, France, etc.

C. *Author Co-occurrence Network Analysis*

A total number of 31 authors and 23 links between the authors were shown in Fig. 1 (c). Because authors belong to the organization, so the cooperation between the authors is similar to that of the institution.

Romano P (Paolo Romano), Sakr S (Sakr Sherif), Guo YK (Yike Guo), Ma K (Ma Kun). The four authors have published the most documents and have a higher frequency of collaboration with other authors.

D. *Paolo Romano*

Dr. Paolo Romano is a senior researcher at the division systems group at INESC-ID, his main research directions are Distributed Data Management, Dependability, Cloud Computing, and Autonomic Computing.

E. *Sakr Sherif*

Sakr Sherif is currently a Professor of Computer Science at King Saud bin Abdulaziz University for Health Sciences. His research interest is data and information management in general, particularly in big data processing systems, big data analytics, data science and big data management in cloud computing platforms. And Dr. Sakr has published more than 100 refereed research publications in international journals and conferences.

F. *Yike Guo*

Yike Guo is an Imperial College London Parallel Computing Center Technical Director, Lifetime Professor in London E-Science Research Center, and the Chairman and CEO of the board of directors of InforSense Ltd. His main research direction is large-scale data mining and parallel computing.

G. *Ma Kun*

Ma Kun, Professor of Key Laboratory of Intelligent Computing Technology for network environment in Shandong Province, member of IEEE, member of China Computer Federation, the main research directions are Big Data Management for Multi-tenant Applications in the Cloud, and Data Intensive Computing.

IV. Co-citation Network Analysis of NoSQL

A. *Author Co-citation Network Analysis*

The higher the frequency of the authors, the stronger the academic authority, so the author have analyzed for authors cited by the data above this paper. Fig. 2 is the Author co-citation map with 180 authors and 549 links generated by CiteSpace and the node named anonymous deleted. Form Fig.2, three core researchers are presented.

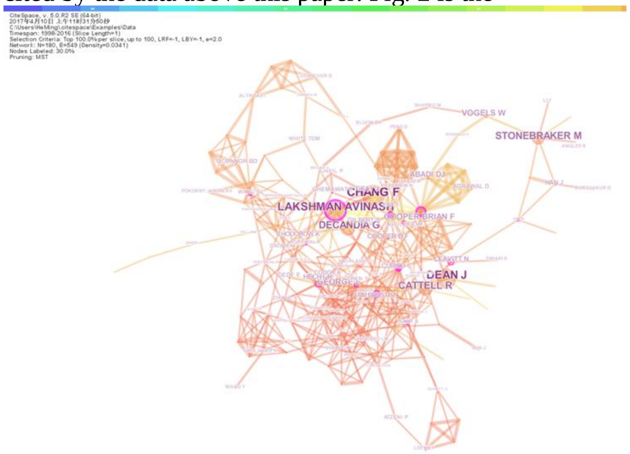

CiteSpace, v. 5.0.R2 SE (64-bit)
2017年4月10日 上午11时31分50秒
C:\Users\HeMing\.citespace\Examples\Data
Timespan: 1998-2016 (Slice Length=1)
Selection Criteria: Top 100.0% per slice, up to 100, LRF=-1, LBY=-1, e=2.0
Network: N=180, E=549 (Density=0.0341)
Nodes Labeled: 30.0%
Pruning: MST
VOGELS W
STONEBRAKER M
ABADI DJ
AGRAWAL D
CHANG F
LAKSHMAN AVINASH
COOPER BRIAN F
DECANDIA G
COOPER B F
GHEMAWAT S
WHITE TOM
OCONNOR BD
HAN J
LEAVITT N
DEAN J
CATTELL R
GEORGE L
HECHT R
ATZENI P
WANG Y

Figure 2 The author co-citation map related to NoSQL researches(1998-2016)

TABLE I. is the key cited authors in the author co-citation map and the cited number more than 10.

B. *Fay Chang*

The largest node is CHANG F (Fay Chang). He is now at Google. In his research, he worked on developing a general, automatic approach to I/O prefetching based on speculative execution. Prior to work at Google, he worked on the Network-attached Secure Disks (NASD) project.

Fay Chang worked with Dean Jeffrey as the first author to complete *Bigtable: a distributed storage system for structured data*.

C. *Jeffrey Dean*

The second largest node is DEAN J (Jeffrey Dean). Jeff Dean is a Google Fellow in the Systems Infrastructure Group. A summa cum laude graduate of the University of Minnesota with a M.S. degree in Computer Science, he obtained a Ph.D. degree in Computer Science from the University of Washington.

Research areas include large–scale distributed systems, performance monitoring, compression techniques, information retrieval, microprocessor architecture, compiler optimizations. Products Jeff has developed for Google include AdSense, MapReduce, BigTable, and Google Translate.

His 6 papers are included in Web of Science Web of Science Core Collection such as *MapReduce: Simplified data processing on large clusters*.

D. *Lakshman Avinash*

Lakshman Avinash is the third largest node with a purple ring, which means that he has a high betweenness centrality and tends to be strategically important in terms of the macroscopic structure of a new work. The reason he was known to most of the people is that he co-invented Amazon Dynamo and invented Apache Cassandra. His main papers include Cassandra: a decentralized structured storage system (cited 297 in Web of Science Core Collection), Dynamo: amazon's highly available key-value store.

Currently Avinash is the CEO and co-founder of Hedvig founded in 2012. Hedvig is positioned as a pure SDS (Software Defined Storage) company to help companies become more responsive to the data demands of today's digital businesses, which received a $ 21.5 million C round of financing on March 4, 2017.

TABLE I. THE KEY CITED AUTHORS IN THE AUTHOR CO-CITATION MAP(THE CITED NUMBER≥10)

No.	Frequency	Centrality	Year	Author
1	34	0.03	2010	CHANG F
2	31	0.08	2013	DEAN J
3	28	0.21	2012	LAKSHMAN AVINASH
4	22	0.01	2012	DECANDIA G
5	21	0.04	2013	CATTELL R
7	20	0.07	2014	STONEBRAKER M
8	10	0.04	2014	GEORGE L
9	10	0.05	2014	VOGELS W

E. *Document Co-citation Analysis*

Fig. 3 shows a document co-citation map with 97 documents and 242 co-citation links. Each node in the graph represents a document in which the thickness of the circle is proportional to the number of citations in the corresponding year.

The core documents shown in Fig. 3 constitutes the most important knowledge foundation in the NoSQL domain. There are 8 core documents cited more than 5 times shown in TABLE II.

Bigtable: A distributed storage system for structured data

The most cited core document is *Bigtable: A distributed storage system for structured data* published in 2008 by Chang Fay as the first author and Jeffrey Dean, and the citation frequency is 18 times. This paper laid the foundation of HBase which was published in 2006 and cited up to 510 times in Web of Science Core Collection. Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers [9]. This article describes the simple data model, dynamic control data layout and format for the client provided by Bigtable, and describes the design and implementation of Bigtable.

F. *MapReduce: Simplified data processing on large clusters*

The second place is MapReduce: Simplified data processing on large clusters published in 2008 by Dean Jean. MapReduce is a programming model and an associated implementation for processing and generating

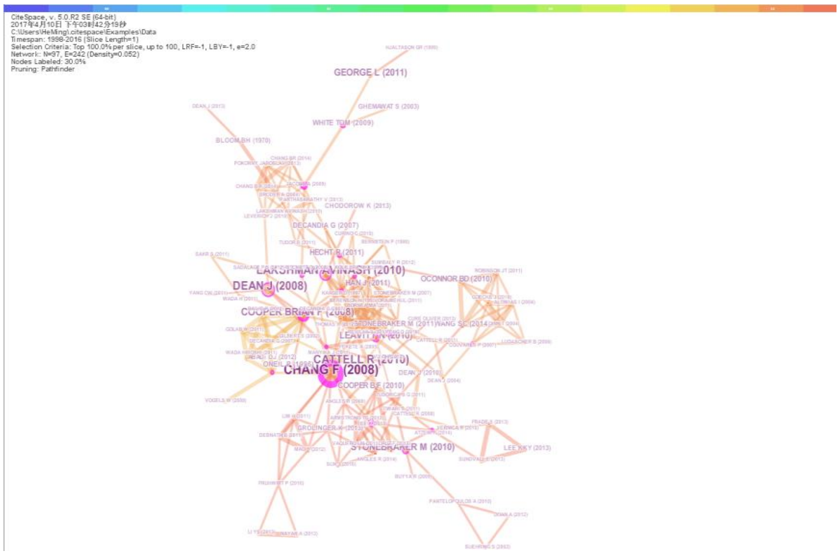


Figure 3 The document co-citation map related to NoSQL researches(1998-2016)

TABLE II. THE KEY CITED AUTHORS IN THE AUTHOR CO-CITATION MAP(THE CITED NUMBER≥5)

No.	Frequency	Centrality	Author	Year	Document
1	18(510)	0.7	CHANG F	2008	Bigtable: A distributed storage system for structured data
2	12(3438)	0.18	DEAN J	2008	Mapreduce: Simplified data processing on large clusters
3	12(168)	0.03	CATTELL R	2010	Scalable SQL and NoSQL Data Stores
4	11(297)	0.17	LAKSHMAN AVINASH	2010	Cassandra: a decentralized structured storage system
5	9(160)	0.32	COOPER BRIAN F	2008	Pnuts: Yahoo!'s hosted data serving platform
6	8(143)	0.04	GEORGE L	2011	Hbase: The Definitive Guide: Random Access to Your Planet-Size Data
7	8(87)	0.12	STONEBRAKER M	2010	SQL Databases v. NoSQL Databases
8	7(90)	0.16	LEAVITT N	2010	Will NoSQL Databases Live Up to Their Promise?

The number in brackets is the cited times in Web of Science Core Collection

large data sets. And many real world tasks are expressible in this model, as shown in this paper [10]. The citation of the article in the Web of Science Core Collection is as high as 3438 times.

V. RESEARCH HOTSPOTS AND FRONTIERS

A. *Research Hotspot Analysis*

Keyword is the core in obtaining the information of an article. Only by accurately grasp the distribution of key words can we better analyze and study the hotspots.

Keywords were set as the network nodes for analysis. By selecting all the documents that appear in each time period, the keywords knowledge mapping was constructed with some irrelevant keywords (such as: lung cancer) deleted. We can get the Fig. 4 with 38 keywords and 58 links.

The largest node in the mapping except NoSQL is nosql database, other larger nodes are big data, cloud computing, mapreduce, mongodb, and so on. Based on Fig. 4, we can find the main areas of the current development states of NoSQL and several important branches of NoSQL domain.

B. *Research Trend and Frontier Analysis*

Research frontier analysis can provide researchers the latest information, and then quickly provide valuable information or references in their potential research area [8]. The development trends and research frontiers can be analyzed according to the keyword frequency changed in the trend. Therefore, the author changed the keyword

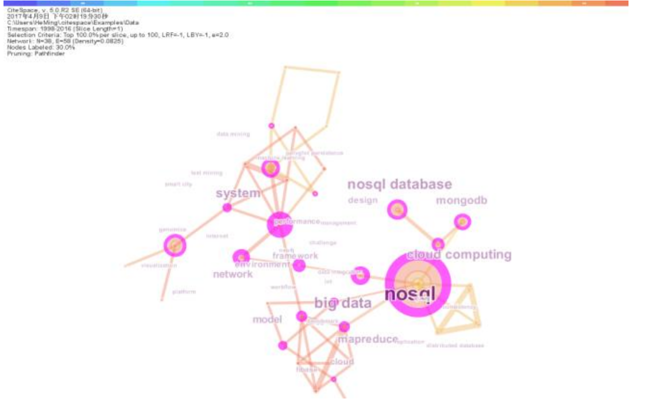


Figure 4 The keyword network related to NoSQL researches (1998-2016)

network to time zone view which provided by CiteSpace, and got Fig. 5.

Research Trends

C. The origin of NoSQL

Although the "data warehouse" in 1998 is the first hotspot in Fig. 7, NoSQL is not originated from the "data warehouse" according to the reference situation of view. It is obvious that NoSQL has evolved from "system", "model", "network", "database" between 2006 and 2007.

1) Development period

In 2009-2010, the concept of opening distributed non-relational database was proposed, but the development of this technology is still at an unexpected stage.

Until 2011, with the rise of "cloud computing" and "MapReduce" technology, NoSQL field has also been unprecedented developing.

In the 2012-2013, NoSQL technology continues to develop forward, and puts some theoretical knowledge into practice, mainly reflected in the nodes named as “NoSQL database”, “Web”, “cloud storage”, and so on.

2) differentiation trend

Since 2014, NoSQL researches gradually split into various fields, mainly focusing on fields such as "bioinformatics", "text mining".

From 2015 to 2016, the differentiation trend is more pronounced. Some main research topics are “mongodb”, “framework”, “cloud”, “hbase”, “data integration”, etc.

Therefore, it can be seen that NoSQL has experienced from the concept to the continuous improvement, and achieved the change from technology theory to practice.

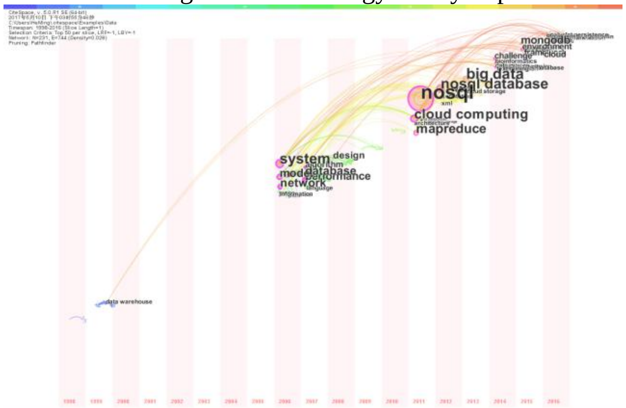


Figure 5 The keyword time zone network related to NoSQL researches (1998-2016)

TABLE III. THE KEYWORDS RELATED TO NOSQL RESEARCHES FROM 2014 TO 2016 (THE FREQUENCY>1)

Frequency	Centrality	Year	Keyword	Frequency	Centrality	Year	Keyword
4	0.13	2014	challenge	2	0.05	2015	association
2	0.11	2014	data mining	2	0.02	2015	workflow
2	0.1	2014	bioinformatics	4	0.05	2016	cloud
2	0.06	2014	availability	3	0.01	2016	hbase
2	0.04	2014	internet	2	0.07	2016	data integration
2	0.04	2014	smart city	2	0.04	2016	neo4j
2	0.04	2014	text mining	2	0.02	2016	benchmark
2	0	2014	consistency	2	0.02	2016	platform
2	0	2014	distributed database	2	0	2016	genomics
2	0	2014	machine learning	2	0	2016	iot
2	0	2014	replication	2	0	2016	management
7	0.01	2015	mongodb	2	0	2016	polyglot persistence
4	0.15	2015	framework	2	0	2016	query
4	0.11	2015	environment	2	0	2016	visualization

And now it is attempting to perfect each branch and utilized by more fields.

Research Frontiers

By the analysis of NoSQL research trend in recent years, the possible frontiers in the future can be found out, such as “data mining”, “data integration”, “and iot”.

The detailed information of keywords which occurrence frequencies higher than 1 from 2014 to 2016 are listed in TABLE III.

VI. CONCLUSION

In the paper, using mapping knowledge domains Software CiteSpace Ⅴ, visualization analysis on 144 documents in NoSQL field were studied to analyze research distribution, co-citation situation as well as the hotspots and frontiers.

A. It Revealed the Distribution Situation of NoSQL Research between Countries, Institutions, and Authors

- The Chinese academy of sciences, Tsinghua University, University of California Berkeley, have published the largest number of documents, leading the

- way in NoSQL researches.

- The United States ranks first in the world in the research field of NoSQL, followed by China, Canada, and France.
- In these 144 documents, Paolo Romano, Sakr Sherif, Yike Guo, Ma Kun have published the most documents and have a higher frequency of collaboration with other authors.

B. It Defined the Key Researchers and Documents

- Fay Chang, Jeffrey Dean, and Lakshman Avinash are three core researchers in which Lakshman Avinash may bring greater influence to NoSQL research in the future because he has a high betweenness centrality.
- Bigtable: A distributed storage system for structured data and MapReduce: Simplified data processing on large clusters laid the foundation for NoSQL researches.
- It showed the hotspots and frontier related to NoSQL ersearches
- The research hotspots are shown out, such as big data, NoSQL database, system cloud, computing, MapReduce, mongodb, etc.
- The technical frontiers of NoSQL like data mining, data integration are discovered.

In the end, this study is summarized: NoSQL research started relatively late, but the development is very rapid. Only in about ten years, it has experienced from reintroduce of concept to continuous improvement, and achieved the change from technology theory to practice with a lot of excellent researchers springing out. There is still huge development space to research and explore in the future.

REFERENCES

[1] Kai Fan. "NoSQL database overview." programmer 6(2010):76-78.

[2] Leavitt, Neal. "Will NoSQL databases live up to their promise?." Computer 43.2 (2010).

[3] Mohan, C. "History repeats itself: sensible and NonsenSQL aspects of the NoSQL hoopla." Proceedings of the 16th International Conference on Extending Database Technology. ACM, 2013.

[4] "Amazon Goes Back to the Future With 'NoSQL' Database."unpublished.

[5] Chen, Chaomei. "Searching for intellectual turning points: Progressive knowledge domain visualization." Proceedings of the National Academy of Sciences 101.suppl 1 (2004): 5303-5310.

[6] Chen, Chaomei. "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature." Journal of the American Society for information Science and Technology 57.3 (2006): 359-377.

[7] Chen, Chaomei, Fidelia Ibekwe‐SanJuan, and Jianhua Hou. "The structure and dynamics of cocitation clusters: A multiple‐perspective cocitation analysis." Journal of the American Society for Information Science and Technology 61.7 (2010): 1386-1409.

[8] Liu, Hailong, et al. "Visualization Analysis of Subject, Region, Author, and Citation on Crop Growth Model by CiteSpace II Software." Knowledge Engineering and Management. Springer Berlin Heidelberg, 2014. 243-252.

[9] Chang, Fay, et al. "Bigtable: A distributed storage system for structured data." ACM Transactions on Computer Systems (TOCS) 26.2 (2008): 4.

[10] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51.1 (2008): 107-113.

Study on Static Task Scheduling Based on Heterogeneous Multi-Core Processor

Shen Yang
Institute of Computer Systems, South China University of Technology
(Guangzhou China 510006)
Shen Yang, born in 1978. PhD. candidate. His research interests focus on computer architecture and system software

Qi Deyu
Institute of Computer Systems, South China University of Technology
(Guangzhou China 510006)
QI Deyu, born in 1959. PhD., professor, Ph.D. supervisor. His research interests include computer architecture, software architecture and computer system security.

***Abstract*—Aiming at the situation of priority scheduling algorithm of chaos, later redundant processing in the task scheduling on current multi-core heterogeneous processors, this paper proposes a task scheduling algorithm with weighted priority algorithm-- WPTS. It is related to three attribute values of the main reference tasks, which can be obtained by weighted comparison that can overcome the confusion and redundancy of task selection to a certain extent. And it can maintain the priority processing of priority scheduling processor in the processor task allocation process to improve the computational efficiency, reduce idle memory accounted rate. In addition, redundant task processing is introduced in order to achieve a better recovery of the processor’ s time period and the effect of time reduction. Compared with the HCPED algorithm and HDEFT algorithm, the WPTS algorithm has a better performance.**

Keywords-Heterogeneous multi-core platform; Static task scheduling; Weighted finite assessment; Time control; Redundant task

I. INTRODUCTION

Because of the high chip utilization rate and the low power consumption, the heterogeneous multi-core processor becomes the important direction of the system in support and development for the multi application platform. But it also brings new challenges to the whole era of computing processor’s development, which means how multi-core processor threads scheduling optimization makes the execution time, execution period or power consumption of processor performance significantly improved. Therefore, as the most important index affecting the integrated performance of the entire multi-core processor, heterogeneous task scheduling becomes a hot research topic in current time. The following is the analysis of three different types of task scheduling algorithm--CPFD/HCPFD/HDEFT, summarizes the evaluation index, and combined with the operating characteristics of heterogeneous multi-core processor, implementing organic combination of three algorithms, weighted parameter values make priority system which realize service mapping in order. This kind of task scheduling algorithm can make up for the deficiency of multi application performance, and has important reference value to the development of the static task scheduling and the technology promotion.

II. ARCHITECTURE OF WPTS ALGORITHM

A. Analysis of Three Kinds of Algorithms

The current general-purpose heterogeneous multi-core processor scheduling algorithm includes CPFD/HDEFT/HCPFD algorithms. Among them, standard of priority classification of the HCPFD algorithm is key work and starting time of the work, one by one to assign the task to processor mapping time, integrated with idle time of system processor to distribute allocate tasks. CPFD algorithm adopts the method of calculating the entrance node to divide the priority, which means the thread distance of the hierarchical from the task node to the entry node that will be scheduled to the corresponding completion time on the processors in batches completing the task in time. With “O”referring data mixed degree in time axis, “V”referring task node inventory. The HDEFT algorithm uses a SUM type priority evaluation mechanism in the algorithm task allocation stage, and in the stage of processor and task link mapping start the complex moment and insert of task.

Comparison of the three algorithms tells adaptive scheduling of CPFD/HCPFD is not very ideal, analysis the reason knows that taking only priority evaluation reference from the algorithm, and does not take into account the existence of the multi constraint in task mapping and influence on scheduling performance of the whole communication overhead for mapping produced. The HDEFT algorithm has no high degree of complexity in time, so the time thread is small. But there is a lack of task after the complex moment of the overflow of redundant processing, but also to extend the overall task scheduling time spent, while taking up more of resources the processor. The following is the summary of the pros and cons of the above algorithm, designing the WPTS algorithm with validation for scheduling algorithm.

B. Execution of WPTS Algorithm

Generally speaking, the WPTS algorithm has two stages of the implementation of the priority decision and the bridge mapping between the task and the multi-core processors.

Among them, as the core of the first stage, the task layer and the weight evaluation are related to the length of the thread of the algorithm, which determines the performance level of the algorithm. In the second stage, the multi task and multi processor allocation is implemented, and the process of the optimization of task scheduling results is derived.

C. Achievement of Principle in WPTS Algorithm

1) Design of Priority Evaluation

First of all, to carry out the combination for task. Follow the same merger rules. It need to optimize the task of independent task, the communication overhead task and task which mapping a longer period of time. Followed by the implementation of DAG depth tree priority search, when the task V only has separate successor nodes, the corresponding only has a precursor node. At the same time, the communication overhead between the two tasks must be greater than the task execution overhead of average processor. After the combined, task is denoted as V*, the sum of the combined expenses is calculated, and the operation and the processing are carried out as a whole in the following task scheduling.

Then, the task layer is implemented. First with the “level” as independent markers for the number of layer in DAG, setting the entrance node of the initial value as 0, the global DAG stratified by stripping down in the order, seperating the maximum number of communication of task nodes and the entrance node , the nominal task named as Level value, so the hierarchical can avoid the omission of global search, and determine the weights for subsequent analysis and calculation of scheduling. The calculation method of Level value follows the following formula:

$$\text{level}(v_i) = \text{Max}(\text{level}(v_j)) + 1,\ v_j \in \text{pred}(v_i)$$

Next, Starting the weight evaluation, calculation process needs to consider the the priority level’s influence of three attributes on task finishing the degree of ranking. WPTS selects the average communication overhead and the mapping time as the task priority evaluation parameters. Through the weight difference, the computation cost of the global server means ACC, the formula refers (a). Communication overhead includes two layer which includes data transfer ADTC and ADRC. Calculation formula refers(b) and (C). X is the successor node parameter, and Y is the precursor node parameter.

$$ACC(v_i) = \sum w_{i,j}/m \tag{α}$$

$$ADTC(v_i) = \frac{1}{x}\sum_{j=1}^{x} C(v_i, v_j) \tag{β}$$

$$ADRC(v_i) = \frac{1}{y}\sum_{k=1}^{y} C(v_k, v_i) \tag{χ}$$

At the same time, weight is defined as the value of the task V, which means the sum of the three tasks (ADTC/ADRC/ACC) that is calculated as follows:

$$\text{weight}(v_i) = \text{ADTC}(v_i) + \text{ADRC}(v_i) + \text{ACC}(v_i)$$

2) Task mapping processing

Process from task to mapping processor includes task mapping and processor redundancy processing

In the process of mapping the entire task, through all the core processors, and the task assigned to the earliest completed idle processor, the job completion time is recorded as EFT1. V allocation and idle period processor task complex moment to complete the node named EFT2. From the moment of the re-engraved precursor node to all processors assigned and in the free time becomes EFT3. Compared with the three aspects, the optimal processing distribution path is selected. The corresponding calculation formula is as follows:

$$\text{EFT1} = \underset{0\leqslant n<|p|}{\text{Min}} \{\text{AST}(v_i, p_n) + w(v_i, p_n)\}$$

$$\text{EFT2} = \underset{0\leqslant n<|p|}{\text{Min}} \left\{ \text{Max}\left\{ \begin{matrix} \text{Avail}(p_n), \\ \text{AFT}(v_i, p_k) + c(v_{par}, v_i) \end{matrix} \right\} + w(v_i, p_n) \right\}$$

$$\text{EFT3} = \underset{0\leqslant n<|p|}{\text{Min}} \left\{ \text{Max}\left\{ \begin{matrix} \text{Avail}(p_n) + w(v_{par}, p_n), \\ \text{AFT}(v_i, p_k) + c(v_{par}, v_i) \end{matrix} \right\} + w(v_i, p_n) \right\}$$

Through the in-depth study of the DAG map, the redundant task processing is the best after the first layer mapping, which means the most close to the original value of the redundant decision after the completion of the adjacent task layer scheduling, and it is easier to accurate correction. Therefore, the step type colloid redundant processing is adopted to carry out correction layer by layer until the redundant item is empty.

III. TIME ANALYSIS OF WPTS ALGORITHM

In the process of task merging, a global DAG layer’s deep analysis is needed to obtain the time complexity O (v+e), which are the number of layers and the number of edges in the graph. Calculating task hierarchical node level value needs to comply with the complexity of the function O (n+e). And in the corresponding task weight evaluation, it needs to make bread search for DAG graph firstly to get the node weight initial value and the entrance node in connection of the key path, the complexity function is O (n2).

The above mentioned functions of time complexity together with the time complexity determines the time length of the task to the processor mapping stage. Therefore, to enhance the processing performance of multi core processor for static multiple task scheduling, it is needed to optimize the DAG layer from the time complexity O. The WPTS algorithm uses O (V + e) + O (n + e) + O (n2) + O (kpm + k2m), which means O (V3) combined with algorithm in suppressing DAG layer complexity at the same time, in order to solve engraved redundancy with reducing redundant scheduling processing time to avoid occupation of processor resource and waste of time.

IV. RESEARCH ON EXPERIMENTAL SCHEDULING MEASUREMENT AND DATA

A. Evaluation of Performance Parameter

In the static task scheduling, the resources and data traffic of overall scheduling is small, so in addition to the total length of the practice of a scheduling algorithm, it needs to add performance indicators. Specific evaluation and reference formulas are as follows

1)Setting “makespan” as the maximum scheduling length index on the global processor.

2)The minimum time for all the critical path tasks seemed as the ratio denominator of scheduling length. SLR looks as algorithm scheduling performance referring association, its value is close to 1, the overall performance of the better.

3) “Efficiency” is set to evaluate the efficiency of the algorithm parameters. The speedup of task scheduling is named as the molecule, and the higher of the value means the higher efficiency of the task scheduling.

The above parameters are calculated as follows:

$$\mathrm{SLR}=\frac{\mathrm{makespan}}{\sum_{t_i\in cp}\min_{p_j\in P}\{W(t_i,p_j)\}}$$

$$\mathrm{Efficiency}=\frac{\mathrm{Speedup}}{|P|}$$

$$\mathrm{Speedup}=\frac{\min_{p_j\in P}\{\sum_{t_i\in DAG}W(t_i,p_j)\}}{\mathrm{makesapn}}$$

B. Data Analysis of Experimental Results

The scheduling performance of the two group WPTS algorithm is tested by simulation experiment in the test group of the same scheduling task.

1)Then a series of task graph happen, according to the DAG task to determine the parameters respectively, then get {30, 40, 50, 60, 70, 80, 90,100}, a (the value of size of DAG is {0.5, 1.0, 2.0}, β (node values {1, 2, 3, 4, 5)}, γ (the value of node{1, 2, 3, 4, 5}), CCR (communication and computation time rate, the value is {0.1, 0.5, 1.0, 5.0, 10.0} generating a total of more than 3000 sets of DAG type, each set has at least 20 different nodes weight type.

2)The task then from the entrance to the global parameters input node processor algorithm, system restores present multi-core processor work by Tesco configuration simulation environment. Join program C to the algorithm to achieve mutual communication, simulation the information interaction and mapping operations of multi-core algorithms. Processing performance is determined by processing efficiency of the same task graph and the length of the processor.

Table I. is detailed program structure and evaluation analysis method referred.

TABLE I . COMPARISON OF ALGORITHM SCHEDULING PERFORMANCE

Algorithm 1	Algorithm 2	Result	Comprehensive
WPTS	HDEFT	Worse:6751 Equal:16523 Better:36726	Ratio:9.47% Worse:17052
	HCPFD	Worse:6884 Equal:11667 Better:41449	Ratio:19.00% Equal:34202
	CPFD	Worse:3417 Equal:6012 Better:50571	Ratio:71.53% Better:128746

It can be seen that in the real simulation environment of random experiment, the three algorithms are in integrated operation, the optimal scheduling ratio of WPTS is the highest, and it is much better than the other three algorithms. With the complexity of scheduling nodes and types of task graphs, the performance advantages of WPTS algorithm and redundant processing rate will become more and more prominent.

V. CONCLUSION

Through the analysis of common task scheduling algorithm for the heterogeneous multi-core processor platform, this paper refers problems and shortcomings to realize the structure optimization and algorithm coordination, and put forward the WPTS scheduling algorithm. The weighted value is used to make the priority of the nominal method, which makes up for the assessment of the single selection of parameter which can be used to accurately reflect the location and attributes of the cluster in the DAG diagram. At the same time, it can bring the task with the shorter thread-the processor mapping, and the elimination of redundant tasks. The optimal scheduling ratio of the new algorithm is as high as 70%, which has a very high and stable processing scheduling performance.

REFERENCE

[1]. Wu Jiajun. Research on task scheduling in multi core and multi thread processor. Anhui: China Academy of Sciences Institute of computing technology, Ph.D. thesis, 2011.4-17

[2] Jiang Yun Lian. Research on task scheduling problem of parallel heterogeneous systems D. Anhui: University of Science & Technology China master's degree thesis, 2012:22-43

New Method of Image Smoothing and Edge Detection Based on Nonlinear Ambiguity Function

Yanhong Jin
College of Mobile Telecommunications Chongqing University of Posts and Telecom
(Chongqing China, 401520)
e-mail: 925411110@qq.com

***Abstract*—With the continuous development of the national economy and science and technology, the scope of the application of images is constantly expanding. The image has many uncertainties, and the improvement of this feature is often blurred and can not be accurately calibrated, which results in complex and diverse processing techniques. Image smoothing and edge detection are very important feature technologies in image processing, and they also the research focus in the field of image processing. This paper studies the method of image smoothing and edge detection based on nonlinear ambiguity function, and discusses the method.**

Keywords- Image; Ambiguity function

I. Introduction

Image is the link between human and information, and it is the visual foundation of the world. The image is affected by many factors in the formation and transfer process which brings a lot of uncertainty on the recognition[1], similar to the fuzzy logic inference of human knowledge on the algorithm that can discover the interference factors in the image data from the image characteristics. From point of view, process information detection image is uncertainty, however, treatment of the uncertainty problem of fuzzy logic shows the effectiveness to a certain extent.

II. Overview of image processing techniques

The image is seemed as a digital signal on the computer, and its forming process will be accompanied by many disturbing factors, image smoothing and image edge processing technique is very important that contains the image and separability of most of the information. The main purpose of image smoothing is to reduce the noise of the image, and the first step is image processing, because most of the image in forms have some noise factors in the process of transmission, so the technology is bound properly to the edge detection of direct impact [2]. Image edge detection refers to the detection algorithm through the collection of gray image edge pixel step change certain, and the image pixels' direction of the shape and edge step of object contour information can be displayed which is an important feature extraction in image recognition[3,4].

Image smoothing and edge detection technology has become the basis for the study of image processing from the image, since the scholars began to study image smoothing and edge detection technology. However, the key technology in initial studies have focused on how to establish the mathematical model, and the image processing technology based on the mathematical model of some targeted with a wide range that was not be covered. In 80s, Pal and King put forward the fuzzy edge algorithmfor image processing technology, the image processing method based on fuzzy technology can effectively separate the object from the background, and applied in the field of medicine and computer pattern recognition, so scholars began to pay attention to application of fuzzy techniques in image processing, and gradually to the the research and development[5].

III. Fuzzy technology

A. *Development of Fuzzy Technology*

In 1960s, American scholar Professor L.A.Zadeh proposed thetheory of fuzzy sets, and in the mathematical basis, fuzzy theory mainly includes the fuzzy set theory, fuzzy logic, fuzzy reasoning and fuzzy control which breaks through the end of nineteenth Century the German mathematician G. Contor founded the classical set theory limitation$_{[4]}$.

The membership function can be expressed as a fuzzy transition process, and the quantitative representation of the concept is established. The mathematical basis of the fuzzy theory is established. P.N.Marinos published a research report on fuzzy logic, which really marks the birth of fuzzy logic. Fuzzy logic is different from classical two valued logic. Fuzzy logic is a continuous logic, anda fuzzy proposition is a sentence that can determine the degree of membership. Its truth value is in any [0 or 1] interval$_{[1]}$. Then, the fuzzy technology for automatic control of the steam engine by a British scholar E.H.Mamdani, and achieved good control effect, it successfully created the people of industrial control direction of fuzzy control, so as to construct the fuzzy control system, the fuzzy control has become a kind of new technology that has been applied in various fields.

Comparatively speaking, fuzzy theory has many advantages. Firstly, fuzzy theory has some theories and methodsto show the natural semantic and makes it into a substance capable of accepting and understanding, in order to improve the cognition and recognition ability of the computer; secondly, the theory of fuzzy reasoning method has fuzzy logic and it is similar to the human brain, thus it makes the computer become more intelligent, such as fuzzy control and industrial products system can reflect these advantages; furthermore, the theory of fuzzy mathematics theory is more widelythan general application, but in various application areas of science and technology and economic development in recent years, fuzzy theory and some optimization algorithms such as neural network, genetic algorithm group a computational intelligence technique. With the development and cross continuously, this theory will promote the human society towards more intelligent society by one step forward.

B. *Overview of Fuzzy Technology*

1) Concept of fuzzy technology

Based on the fuzzy *set* theory, the set of elements is set up by [0, 1], and real interval representation are used, and “1” means logical truth, and “0” means logical false. However, when fuzzy logic is used to establish the set of set elements, the value of the proposition is true except that it is “1” and falseis “0”, so that any value between 0~1 can be taken, such as 0.6, which means the proposition is true or false to a certain extent. The fuzzy concept can be described by fuzzy set, and the function expressed by the set is the membership function:

$$\mu : X \to [0, 1]$$

The collection described by the membership function μ called fuzzy set to a fuzzy subset A conclusion on field X, μA(x) means the membership of set A, and it is also expressed as a mapping on the domain:

$$\mu_A : X \to [0, 1]$$

$$X \mapsto \mu_A(X)$$

As shown in Fig. 1, the general collection can define a clear boundary, such as under the age of 30 is young, however in the fuzzy set, it cannot define a clear boundary for young people under 30 years of age or 40 years old, which means 40 years old also belongs to the youth. But they belong to different levels of youth.

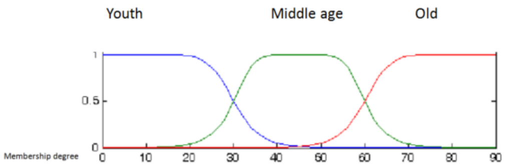


Figure 1. Typical subordinate functions of language values “Youth”, “middle age” and “old”

2) Operation of fuzzy sets

The three fuzzy sets A, B and C on the domain X:

a) If for $\forall x \in X$,then $A(x) = B(x)$ which means $A = B$:

$$A = B \Leftrightarrow A(x) = B(x)(\forall x \in X)$$

b) If for $\forall x \in X$,then $A(x) \le B(x)$:

$$B \Leftrightarrow A(x) \le B(x)(\forall x \in X)$$

c) If for $\forall x \in X$,then $B(x) = 1 - A(x)$, which means B is the remainder of A.

d)If for $\forall x \in X$,then $C(x) = max\{A(x), B(x)\}$ or $C(x) = A(x) \cup B(x)$, which means C is union set of A and B.

$$C = A \cup B \Leftrightarrow C(x) = A(x) \cup B(x)$$

5) If for $\forall x \in X$,then $C(x) = min\{A(x), B(x)\}$ or $C(x) = A(x) \cup B(x)$,which means C is intersection of A and B:

$$C = A \cap B \Leftrightarrow C(x) = A(x) \cap B(x)$$

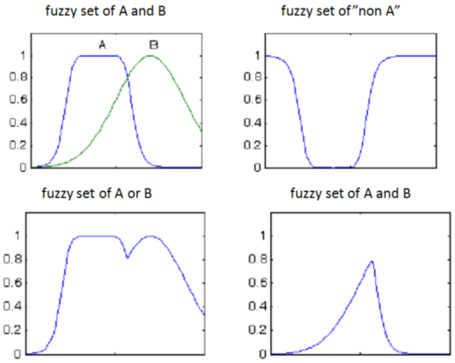


Figure 2. Fuzzy set operation

3) Fuzzy logic and fuzzy language

About quadratic element logical values are usually in {0 and 1}, which are not true or false. However, in many practical problems, it is difficult to make such a true and false judgment. Based on the fuzzy logic value function, no exact value proposition for dividing the determination which is defined to a certain extent, the ambiguity function of the proposition is judged only on closed interval [0,1], when determining the extent of the closer to 1, that proposition is true to the ratio that is higher, and vice versa proposition for the higher percentage is false.

Vague language is usually the simplest "If... Then"rule, such as "if x belongs to A,and Y belongs to B". The compound means"if m is A and X is B, then y is C, or Z is D"; in which, A, B, C, D respectively on domain M, X, Y, Z including semantic fuzzy set values, "If" is part condition, and"then" is the condition.

For the image processing technology in the edge detection, when using the step edge detection edge slope, the parameters in the model are difficult to determine, in this situation, the human knowledge application of edge detection algorithm shows a flexibility based on fuzzy theory which is applied to the fuzzy logic expert system thought with using"If... Then" rules to represent human knowledge.

IV. Application of fuzzy technology in image processing technology

In the actual application process, the fuzzy function has certain steps of image smoothing and edge detection based on image technology: firstly, linear and nonlinear fuzzy membership function can replace the traditional definition of correlation based on fuzzy membership function correlation, improved traditional image smoothing algorithm making the noise out of processed images, and it is more clear; secondly, the mode conversion of the input image, which means the image is converted into gray image, its purpose is to reduce the amount of calculation and speed up the computation time for feature extraction; then, the image data began fuzzy processing, and image data is transformed into fuzzy domain data; fuzzy image's pre-processing domain means fuzzy enhancement of image that reduces the image fuzzy uncertainty in the decision; and then determine of the bandwidth and fuzzy threshold with the use of "If... Then" fuzzy language rule determines the degree of its collection, detects the edge data in the fuzzy domain, and finally outputs the edge image in the spatial domain.

Image processing technology is a kind of digital processing method, these methods are to improve the appearance of the image, because of the equipment, channel and objective condition limit, it causes the conversion between the actual scene and image information of the deviation, affecting the image quality of output. In a word, the ambiguity function of the image processing technology can make the image observation and judgment which is more suitable for human eyes based on analysis and machine processing that has important significance to improve the image quality and improve the visual effect.

Reference

[1] Xiaozhuan Zhang, Dongyan Fan.Research on fuzzy feature recognition of edge of inclined license plate[J]. Computer Simulation, 2017, 34(1): 372-375.

[2] Jie Dou, Weiqiang Han, Jingneng Fu.Motion deblurring based on prior model of edge and gradient distribution[J].Electronic Design Engineering, 2017 (6): 66-70.

[3] Qingju Tang, Junyan Liu, Yang Wang,Infrared image edge recognition and defect quantitative detection based on fuzzy C mean clustering and Canny operator[J].Infrared and Laser Engineering, 2016, 45(9): 928001-0928001 (5).

[4] Lingyi Song, Tao Shu, Derong Zhou.Application of Canny edge detection based on super paste set in Pepper image[J].Journal of Industrial and Commercial University Of Chongqing: Natural Science Edition, 2016, 33(3): 38-42.

[5] Weihua Liu.Image edge detection based on fuzzy enhancement and least squares support vector machines[J].Automation and Instrumentation, 2016 (7): 224-225.

2017 International Conference on Computer Network, Electronic and Automation (ICCNEA 2017)

Research and Application of an Intelligent Decision Support System

Xiaoqing Zhou
Center of Computing
China West Normal University
Nanchong, China
e-mail:369656820@qq.com

Jianqiong Xiao
Center of Computing
China West Normal University
Nanchong, China
e-mail:2439951778 @qq.com

Zhiyong Zhou
Center of Computing
China West Normal University
Nanchong, China
e-mail:449735106@qq.com

Jiaxiu Sun
Colleges of Business
China West Normal University
Nanchong, China
e-mail:Zhousun123@163.com

Abstract—This paper discusses a new decision-support system that integrates data warehouse, knowledge warehouse and model warehouse. Contrast to the fixed model of the old decision-support system and its limited application, the new system can overcome the shortcoming of the old system efficiently, and also it can simplify model-obtaining and coding. So the new system strengthens the effectiveness, intelligence and efficiency of the decision.

Keywords-Decision-Support System; Data Mining; Knowledge Discovery; Model Warehouse

I. INTRODUCTION

Although DSS (Decision-Support System) can supply timely, accurate and scientific information, the most advanced SDSS (Spatial Decision Support System) has defects. SDSS integrates the traditional and the new DSS (including data warehouse, OLAP (On-Line Analysis Processing), data mining, data base, and ES), so it can solve many questions. But due to the fixed model of the model warehouse, which cannot adjust according to the change of the condition parameter, the application of the SDSS is limited. So the paper is tries to introduce a new decision-system system that is based on the data warehouse, knowledge warehouse and model warehouse. The new system can update the knowledge of the knowledge warehouse freely by using knowledge warehouse and date warehouse. And also the new system can strengthen the effectiveness, intelligence and efficiency of the decision by the management of the MWMS and the study of the system.

II. MAIN MODULES

The system is composed of model warehouse, knowledge warehouse, method base, data warehouse, OLAP and data mining modules.

A. Model Warehouse

Model warehouse has the following functions: management with classification, memory the necessary model (including using date-mining model) and comprehensive model parameter (in order to choose out the proper model). Machine-detecting technology integrated the artificial intelligence (AI) can accomplish model creating by computer by simulating the data of the date warehouse/ database. The Self-study algorithm by the nerve network of the model-study can adjust the model fine and update the parameter to get the optimal practical model, so the model can keep in chorus with the fact. Flexible software development technology integrated Software Engineering supports model-coding. Model management system is to manage model of the model system and to call/operate model.

B. Knowledge Warehouse

Knowledge Warehouse has the function of obtaining, clearing/transforming/coding, organizing, memorizing, adjusting, and propagating knowledge. KW can accomplish the function by expanding system structure of the date warehouse. KW is composed of six components. ① knowledge/date-obtaining module. It is to switch recessive knowledge to dominant knowledge, which is to say to get recessive knowledge from decision-maker. ②Two feedback loops. One is between knowledge-obtaining module and knowledge-memorizing module. The other is between Extract-Transform-Load (ETL) module and mutual management module. And it is to memorizing the knowledge, which has been verified by the system and to update knowledge warehouse timely. ③ETL module. It is similar with the corresponding module of data warehouse. ④ Knowledge warehouse module. One of the main components is Knowledge Base Management System, which accomplishes the analysis both by knowledge warehouse and model warehouse. ⑤Analysis worktable. It is composed of task controlling, conclusion getting and technology-Management modules. ⑥Interface module. It is to handle the interaction between KBMS and user's interface. The knowledge warehouse system is as following figure 1.

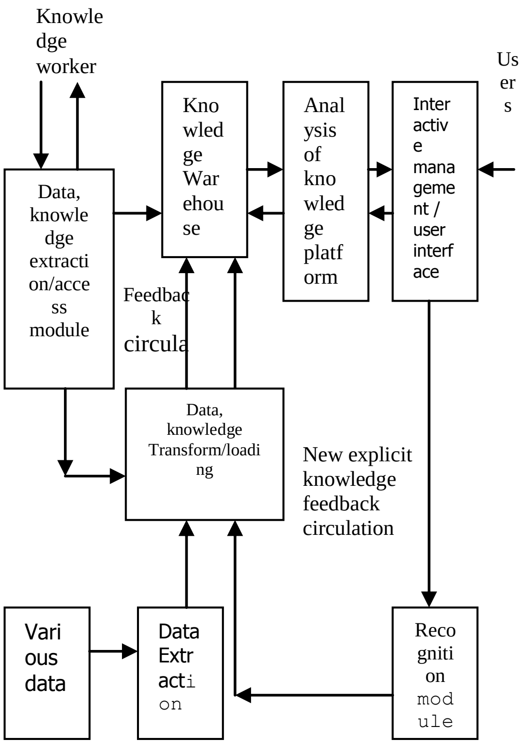

Knowledge worker
Data, knowledge extraction/access module
Knowledge Warehouse
Analysis of knowledge platform
Interactive management / user interface
Users
Feedback circula
Data, knowledge Transform/loading
New explicit knowledge feedback circulation
Various data
Data Extraction
Recognition module

Figure 1. Knowledge warehouse system

C. Method base

Method should be based on the model and be adjusted according to the model in order to calculate. But one model can have several methods. Method base is to supply method for DSS problem model to calculate. And method base management system is to add, delete, revise and search method and to give service for model solving.

D. Data warehouse and OLAP

OLAP is one kind of data warehouse application and it is based on data warehouse. So it can provide decision-makers with analysis results by analyzing and handling. Data warehouse organizes data according to function requirement, the use and granularity of DSS. The key point of OLAP is how to organize data to satisfy user's multi-dimension data analysis.

E. Data-mining

Data-mining module is to mine data to get the needed knowledge according to the model, method and knowledge provided by relevant warehouse. And the result of data mining can be used as new knowledge and model to solid knowledge warehouse and model warehouse.

F. Problem solving and interactive system

Problem solving module is to solve problem by using knowledge, model, method and knowledge of relevant warehouse. Non-structure problem, which cannot be structured, may be solved by deduction system.

III. THE FRAMEWORK AND STRUCTURE OF NEW DECISION-SUPPLY SYSTEM

Figure 2 is the structure of DSS, which integrated DW, KW, MW, MB, OLAP, Data-mining and Problem Solving system. Data mining, knowledge-deduction centers, model-creating units of model warehouse are the intelligence center of DSS which strengthens intelligence property of DSS. And problem solving and interactive system are the function center.

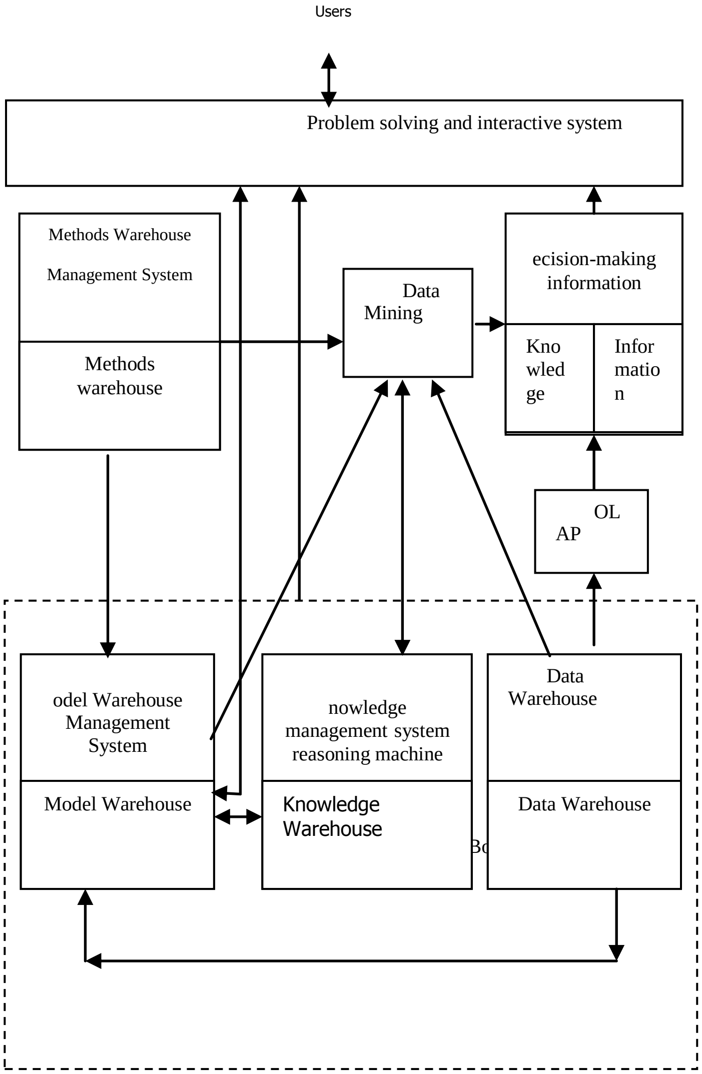

Users
Problem solving and interactive system
Methods Warehouse Management System
Methods warehouse
Data Mining
ecision-making information
Knowledge
Information
OLAP
odel Warehouse Management System
Model Warehouse
nowledge management system reasoning machine
Knowledge Warehouse
Data Warehouse
Data Warehouse
Bo

Figure 2. Integrated DW, KW, the MW DSS system structural drawing

DSS comprises three main parts. The first one is the integration of MW system, DS system and DW system. And it is the basis of decision-support system to provide assist decision-making information of Quantitative analysis (Model Calculation). The second one includes DW and OLAP, which extract spatial data and information from DW. The third one is the integration of experts system and data mining system. Data mining mines knowledge from DB and DW and puts it into knowledge warehouse of experts system, then experts system analyzes. The three parts are integrated. Users can choose one part for decision, either two or three according to the fact. The traditional DSS chooses the first part, IDSS chooses the first part and the third part, and the new DSS chooses the second part and data mining of the third part. The new DSS integrates the three parts by using problem solving and interactive system can give better assist decision-making decision.

Generally speaking, three integrated parts; three warehouses and the application of closed cycle free back and the introduction of MW system is the characteristic of the framework, which makes it more intelligent.

IV. KEY TECHNOLOGY TO ACCOMPLISH THE DECISION-SUPPORT SYSTEM

A. Data mining and text mining

Data mining is to find out unobvious pattern and to acquire needed knowledge in order to help enterprises make decision more scientific and more accurate by analyzing and handling large number of data. Text mining is to acquire valuable information from all kinds of text information. The text source can be Web, fax, E-mail, example and other kinds of text. The decision-makers can extract useful information according to rules and guides who have been defined advanced to make decision.

B. Modeling

Workers should define objective function, decision variables and its' weight. They also should make definite restrictive conditions and coefficients of variables according to the decision variables. So the model that comprises the elements (decision variables, coefficients, restrictive conditions and objective function) can reflect the invisible knowledge obviously.

C. Decision-support tools

It is the tool which uses existed knowledge to help make decision. It needs much technology and tools, including AI, expert system, software engineering, knowledge search tool, knowledge explaining tool and multi-dimensional tool and so on.

D. Intelligence-support technology

it includes: ① model warehouse system should be designed to accomplish its function. ②Interface: All parts are joined by interface. Model, data and knowledge are separated parts that should be integrated. So the interface is very important. Interface should have the function of saving and extracting data, calling and operating the model, and knowledge reasoning. ③system integrated: an integrated system should integrate all parts by words according to the fact.

V. EXAMPLES

Here is an example for a domestic large-scale machinery limited corporation. The old quality management decision-support system of the corporation cannot adjust effectively due to the fixed model coefficients and cause low efficiency. In order to increase production the decision must be changed. Generally, quality breakdown and quality cost will be the core after checking relevant files and investigating the decision-makers. And the analysis should be started from suppliers, manufactories, employees, products and time.

As for products, they should be analyzed by one product or product classification. But one unit can analyze manufactories, employees and suppliers. Time itself is a dimensional data. The quality analysis for every employee, product, supplier and manufactory can be made yearly, quarterly and monthly and the result (graph or table) can help make decision. But if the model coefficients cannot adjust finely, the above result cannot be obtained. The new decision-support system that integrated KW, MW, and DW can be developed and it is based on Windows2000, SQLServer2000, and Excel 2000. The new system can revise model coefficients automatically according to the change of time, employee, product and manufactory. It also can call method of method base and handle by OLAP; the results are as figure 3. So decision-makers will know the worse part and can make better decision by getting the right reason.

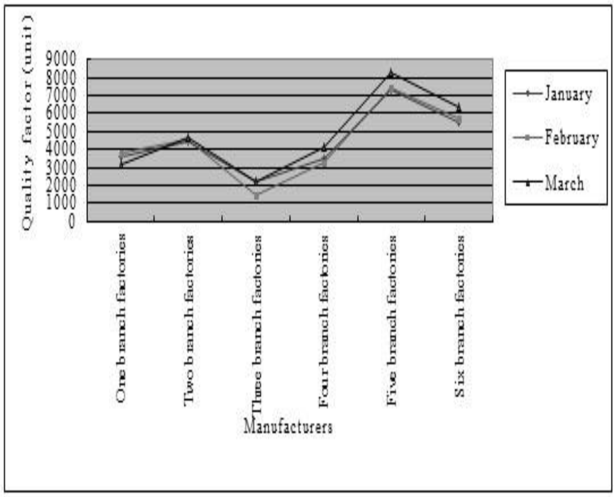


Figure 3. Product qualities pursues every branch factory comparatively

REFERENCES

[1] Chen Song-can, ZhuYu-lian, Zhang-Daoqiang, eta1.Feature Extraction Approaches Based on Matrix PaRem:MatPCA and MatFLDA[J]. Pattern Recognition Legers, 2005, 26(8):l157 一 l167.

[2] RAO Yi-ning,LIU Qiang,DU Xiao-li,YE Peng,Research and Design of Extensible Knowledge Database Model Applied to Intelligent Chinese Search Engine [J],Application Research of Computers 2006,23(6):223-226.

[3] ZHAO Han,DONG Xiao-hui,FENG Bao-lin,WU Zhao-yun,Modeling and Application on Decision Support System Based on Knowledge Warehouse [J],Journal of Systems & Management,2008,17(3) : 327-331.

[4] Feng Qing, Yu Suihuai, Yang Yanpu, Product DSS Model Based on Cloud Service [J], China Mechanical Engineering, 2016, 24(15):201-20159.

[5] Yang Fenfen,Wang Ying, The Research and Design of the Decision Support System about Agr ciultural Machinery [J], Journal of Agricultural Mechanization Research, 2014, (3) :35-38.

[6] LIU Bo-yuan, FAN Wen-hui, XIAO Tian-yuan, Development of Decision Support System, Journal of System Simulation, 2011, 23(7):241-244.

[7] Xu Wei Based on Knowledge Discovery Mechanism of Enterprise Decision Support Systems Research [D], 2013, 12.

[8] YE Zong-qiang; TONG Xin-shun; QIN Xiao-kang. The Resarch on Tobacco Logistics Decision Support System under the Background of Big Data. Logistics Engineering and Management, ,2015,37（1） : 113-116

[9] XU Gang-qiang;LIN Yan.Applying to College Aided Decision Support System Based on Data Mining, 2014,33(4):106-109

[10] SONG Fu-Gen; WANG Yi-Sha. Production Decision Support System Based on Cloud Computing. Computer Systems & Applications, 2015,24(4):44-50

Learning Better Classification-based Reordering Model for Phrase-based Translation

Li Fuxue
NiuTrans Lab School of Computer Science and Engineering,
Northeastern University, China,
YingKou Institute of technology,
e-mail:lifuxue119@163.com

Zhu Jingbo
NiuTrans Lab School of Computer Science
and Engineering, Northeastern University, China

Xiao Tong
NiuTrans Lab School of Computer Science
and Engineering, Northeastern University, China

Abstract—**Reordering is of a challenging issue in phrase-based statistical machine translation systems. This paper proposed three techniques to optimize classification-based reordering models for phrase-based translation under the bracket transduction grammar framework. First, a forced decoding technique is adopted to learn reordering samples for maximum entropy model training. Secondly, additional features are learned from the context of two consecutive phrases to enhance the prediction ability of the reordering classifier. Thirdly, the reordering model score is integrated as two feature functions (STRAIGHT and INVERTED) into the log-linear model to improve its discriminative ability. Experimental result demonstrates significant improvements over the baseline in two translation tasks such as Chinese to English and Chinese to Japanese translation.**

Keywords-statistical machine translation; word reordering; log linear model; feature selection

I. INTRODUCTION

The phrase-based translation approach has been a popular and widely used strategy to the statistical machine translation (SMT). In phrase-based statistic machine translation (PBMT), reordering is of a big challenge and a great importance issue, and it is typically handled by two different models such as distortion model and lexicalized reordering model. Distortion models consider the distance of the words or phrases movement (Brown et al., 1993; Koehn et al. 2003). Lexicalized reordering models are proposed to learn phrase orientation base on content (Tillmann, 2004; Koehn et al., 2005; Nagata et al., 2006). In this paper, we focus on lexicalized reordering models for phrase-based translation. Among the lexicalized reordering models, Bracket Transduction Grammar (BTG) restriction is widely used for reordering in SMT (Zens et al., 2004) due to its good tradeoff between efficiency and expressiveness. Under framework of BTG, the reordering task is considered as classification problem and achieves good performance (Abdullah et al., 2014), referred to as the classification-based reordering model (CRM). The maximum entropy classifier is widely adopted by many researchers to implement the CRM (Zens and Ney, 2006; Xiong et al., 2006; Nguyen et al., 2009; Xiang et al., 2011), and is also considered in this work.

In principle, three key issues should be addressed to build effective classification-based reordering models. The first key issue is how to learn reordering samples from bilingual corpus to train the classifier. The traditional way is to learn reordering samples from each sentence pair based on its word alignments. However, it is sensitive to word alignment noise because a word alignment error would result in some incorrect reordering samples and block some desirble reordering samples. To alleviate this problem, this paper presents a forced decoding based approach to learning reordering samples from derivations of each sentence pair instead of word alignments. Secondly, to build a powerful classifier for CRM, e.g. based on maximum entropy model, traditional methods learn classification features only from source and target sides of two consecutive phrases for reordering, e.g., boundary information of both phrases. Since the source-side context of two consecutive phrases can provide more valuable information for reordering, in our work some additional features are learned from the context of two consecutive phrases to enhance the prediction ability of the reordering classifier. Thirdly, reordering model score is typically integrated as one feature function into the log-linear model. Our method considers reordering model score as two feature functions (STRAIGHT and INVERTED) to improve reordering discriminative ability. Experimental results show significant improvements over the baseline in two translation tasks such as Chinese to English and Chinese to Japanese translation.

II. RELATED WORK

A number of approaches have been proposed to address the reordering issue in phrase-based translation. In principle, the reordering approaches can be divided into two categories: pre-reordering and reordering mode at decoding time.

The first category reorders the source language in a preprocessing step before decoding (Nieben and Ney 2001;Collins et al., 2005; Isozaki et al., 2010), this kind of methods aim at arranging source words in a target-like order before decoding. This paper focuses on the reordering model at decoding time.

The second category estimates phrase movement with reordering models at decoding time. In distortion models, IBM models 1 and 2 define the distortion parameters in accordance with the word positions in the sentence pair instead of actual words at those positions (Brown et al., 1993). Models 4 and 5 limit this by replacing absolute word positions with relative word positions (Brown et al., 1993). Lexicalized reordering models introduce reordering probabilities conditioned on the words of each phrase pair, and they distinguish three orientations with respect to the previous phrase pair (Tillmann, 2004; Koehn et al., 2005; Nagata et al., 2006).

Tillman (2004) considers the position of each phrase as a class, and Koehn et al. (2005) extend the classes to any arbitrary number. Galley and Manning (2008) extended the lexicalized reordering model to tackle long-distance reordering. These reordering models learn local orientations with probabilities for each bilingual phrase from training data. However, since reordering is related to concrete phrases, the data sparseness problem may be introduced. Under the restriction of BTG, some researchers had posed the phrase movement problem as a classification problem. Zens and Ney (2006) introduced a maximum entropy classifier for phrase reordering. Xiong et al. (2006) proposed a maximum entropy model to predicate reordering of neighbour blocks (i.e. phrase pairs), and considered straight or inverted orientations. Nguyen et al. (2009) applied a maximum entropy model to learn orientations identified by the hierarchical reordering model. Xiang et al. (2011) introduced a smoothed prior probability to maximum entropy model, and used multiple features based on syntactic parsing to improve reordering in PBMT. Alrajeh and Niranjan (2014) posed phrase movements as a classification problem, and explored a generative learning approach named Bayesian naive Bayes to dealing with phrase reordering. Recently, neural reordering model (Li P et al., 2014) is also adopted to deal the reordering issue and it could address the data sparseness problem.

III. Classification-based Reordering Model for PBMT

Phrase-based SMT systems move from using words as translation units to using phrases, it has been widely used and achieves the state-of-the-art performance. However, reordering is still a crucial issue for PBMT. Many researchers proposed lexicalized reordering models to address this issue (Tillmann, 2004; Koehn et al., 2005; Nagata et al., 2006). In prinpicle, lexicalized reordering models learn local orientations with probabilities for each bilingual phrase from training data. To alleviate the data sparness problem of lexicaliezd reordering, a kind of models which treat the reordering issue as classification problem are proposed under the BTG framework (Zens et al., 2004).

BTG is employed firstly in statistical machine translation in (Wu, 1996). Under the framework of BTG, three rules are adopted to generate the translations:

(1) A→ [A1, A2];

(2) A→＜A1, A2＞;

(3) A→ x / y;

where rule (1) merges two consecutive blocks into a larger blocks in the straight order, rule (2) does the same work in the inverted order and rule (3) translates phrase y into target phrase x and generates a block A.

The maximum entropy-based approach (so called MaxEnt) is widely used to implement classification-based lexicalized reordering models by many researchers (Zens and Ney, 2006; Xiong et al., 2006; Nguyen et al., 2009; Xiang et al., 2011), which is defined as:

$$\Omega = f(o, A_1, A_2) \qquad (1)$$

$$o \in (straight, invered) \qquad (2)$$

$$\Omega = p_\theta(o \mid A_1, A_2) = \frac{\exp(\sum_i \theta_i h_i(o, A_1, A_2))}{\sum_o \exp(\sum_i \theta_i h_i(o, A_1, A_2))} \qquad (3)$$

where o∈(straight, inverted) indicates phrase orientation, h ∈{0,1}is the ith classification feature and θ i is weight of the ith feature.

IV. Learning Better CRM

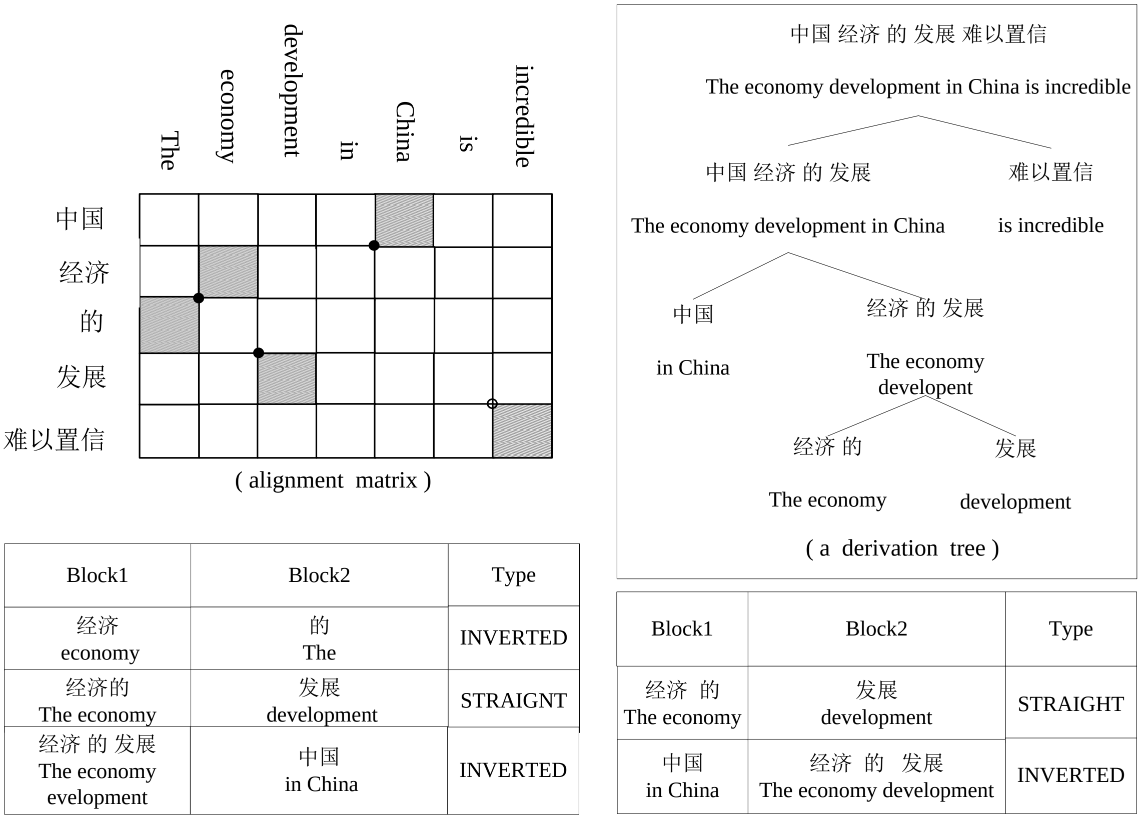
This section presents three optimization techniques to improve classification-based reordering models for PBMT, involving reordering sample generation for training, feature selection for classification and reordering feature functions for decoding. We will discuss these techniques in details as follows.

A. *Reordering Sample Generation for Training*

The first step is to learn reordering samples to train the MaxEnt classifier used by CRM. In traditional method, the reodering samples are learned from bilingual sentence pairs based on word alignments. Given a bilingual sentence pair with its word alignments, we can get the alignment matrix as shown in Figure 1. There are some vertexes shared between two blocks which have four directions: top-left, top-right, bottom-left and bottom-right. The top-right and bottom-left link blocks with the straight order, so we call them INVERTED links. Similarly, we call the top-left and bottom-right STRAIGHT links since they link blocks with the inverted order. For example, in Figure 1, the order of “经济-economy” (Block1) and “的-the” (Block2) is INVERTED, and the order of “经济 的-the economy” and “发展-development” is STRAIGHT. Actually the traditional approach is sensitive to the word alignments, because word alignments errors would result in incorrect reordering training samples and block some desirable reordering samples extracted. For example, the word alignment error [“的”-“the”] introduces some incorrect reordering samples, e.g., {“经济-Economy”, “的-the”, INVERTED}.

To alleviate this problem, this paper adopts a phrase-based forced decoding approach to learning reordering samples from derivation tree (or forest) of each bilingual sentence pair, as shown in the right side of Figure 1. The phrase-based forced decoding technique is different from the typical phrase-based decoding method, in which the derivation of each translation hypothesis must yield the same target sentence during the phrase-based decoding process. In other words, a derivation hypothesis different from the given target sentence could not survive during the phrase-based forced decoding process.

In the CYK decoding process, the words in segmented source sentence are treated as the basic unit, referred to as cell. For each cell that spans from i to j on the source side, the derivations in cell (i, j) was generated by merging derivations from any two neighbouring sub-cells. For each cell (i, j), k is defined as i < k < j. There would be two sub-cells: cell(i, k) and cell (k, j), we can combine the two cells by the straight and inverted rules, and the application of two rules will generate new translation hypotheses, then we drop the derivations which are not yield the target sentence. When the whole source sentence is covered, the decoding process is finished, we can trace back the path of the derivation to learn the details of how to derive the target sentence (translation reference).



Block1	Block2	Type
经济 economy	的 The	INVERTED
经济的 The economy	发展 development	STRAIGNT
经济 的 发展 The economy evelopment	中国 in China	INVERTED

Reordering samples base on word alignments

Block1	Block2	Type
经济 的 The economy	发展 development	STRAIGHT
中国 in China	经济 的 发展 The economy development	INVERTED

Reordering samples base on forced decoding

Figure 1. Alignment matrix and parts of the reordering samples base on word alignments

To alleviate this problem, this paper adopts a phrase-based forced decoding approach to learning reordering samples from derivation tree (or forest) of each bilingual sentence pair, as shown in the right side of Figure 1. The phrase-based forced decoding technique is different from the typical phrase-based decoding method, in which the derivation of each translation hypothesis must yield the same target sentence during the phrase-based decoding process. In other words, a derivation hypothesis different from the given target sentence could not survive during the phrase-based forced decoding process. In the CYK decoding process, the words in segmented source sentence are treated as the basic unit, referred to as cell. For each cell that spans from i to j on the source side, the derivations in cell (i, j) was generated by merging derivations from any two neighbouring sub-cells. For each cell (i, j), k is defined as i < k < j. There would be two sub-cells: cell(i, k) and cell (k, j), we can combine the two cells by the straight and inverted rules, and the application of two rules will generate new translation hypotheses, then we drop the derivations which are not yield the target sentence. When the whole source sentence is covered, the decoding process is finished, we can trace back the path of the derivation to learn the details of how to derive the target sentence (translation reference).

The process of a source sentence which is decoded successfully by forced decoding will form a tree structure, referred to as the derivation tree, as shown Figure 1. The phrase in the node which has two chilidren nodes can be composed by the combination of two phrases in its children nodes. The algorithm of learning reordering samples based on forced decoding is summarized into six steps as follows:

1. Extract translation rules needed for a specific phrase-based SMT paradigm M from bilingual training corpus C;

2. Perform minimum error rate training (MERT) on a development data set to obtain a set of optimized feature weights;

3. For each {s,t}∈C, translate s into accurate t based on M with translation rules learned in step 1 and feature weights optimized in step 2;

4. For each {s,t}∈C, save the derivation forest produced in step 3 as TreeSet.

5. For each derivation tree belongs to {s,t}, Traversing Treei produced in step4 and extracting the reordering samples from the combination of two phrases in children node.

6. Combine the reordering samples belongs to each sentence pair {s,t}, and remove the duplicate reordering samples.

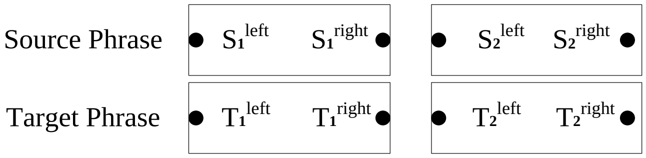
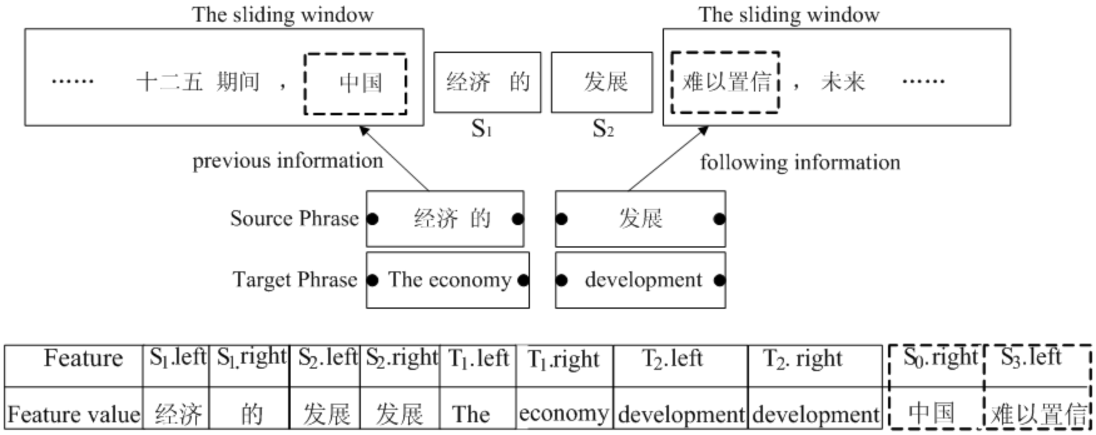

Source Phrase
● S_1^{left} S_1^{right} ●
● S_2^{left} S_2^{right} ●
Target Phrase
● T_1^{left} T_1^{right} ●
● T_2^{left} T_2^{right} ●

Figure 2. Boundary words (black dots) in the two neighboring phrases



Feature	S_1.left	S_1.right	S_2.left	S_2.right	T_1.left	T_1.right	T_2.left	T_2. right	S_0.right	S_3.left
Feature value	经济	的	发展	发展	The	economy	development	development	中国	难以置信

Figure 3. Boundary features (the solid frame) and contextual features (the dotted frame) for the classifier when setting the sliding window K = 1

Take the derivation tree shown in Figure 1 as example, the node with phrase “中国 经济 的 发展—The economy development in China ” can be generated by the combination of “中国—in China” and “经济 的 发展—The economy development” in the inverse order, and we can learn the reordering samples from this combination. In other words, the forced decoding based method learns the reordering samples from the combination of the two phrases, which represents the details of how to generate the derivation. Therefore the quality of reordering samples is much higher than that of traditional methods. In fact, there may be multiple ways to decode a source sentence to target reference by forced decoding technique. In other words, there are several ways to derive the generating tree, referred to as the derivation forest.

Figure 1 shows a derivation tree of the generating forest and parts of reordering samples extracted from the generating forest. From the reordering samples extracted by two methods, we discover that the incorrect reordering samples extracted base on word alignments is discarded in this forced decoding based approach.

B. *Classification Features*

In traditional classification-based reordering model, the maximum entropy classifier generally considers phrase boundary words of reordering examples as features. It can be illustrated as shown in Figure 2. Sleft imeans the most left word in source phrase Si and Sright imeans the most right word in source phrase Si; Tleft imeans the most left word in target phrase Ti and Tright imeans the most right word in Ti. Figure 2 shows the eight boundary words (bold dots) of two consecutive phrases {S1, S2} and their corresponding target phrases {T1, T2}. Boundary words of the source phrases and target phrases are selected as eight features to build the classifier. As shown in Figure 3, these eight features are listed within solid frames. Since a target phrase T2 only contains one word “development”, two boundary word features (T2.left and T2.right) of T2 are the same “development”. In other words, the left-most word is the same as the right-most word, and the rule is also applied to source phrases.

In traditional method, only boundary information (i.e., in the form of eight features) is considered. In our opinion, the source-side context of both two consecutive phrase pairs in the source sentence can also provide more valuable information for reordering. Therefore, in our approach, the contextual information in source sentence is considered to predict the order of two consecutive phrases. Along this line of thinking, we can choose the contextual of the source phrases as additional features. First, the sliding window K is defined as the phrase number that we extend in two directions from the current phrase in source sentence, theoretically the max value of K is can be set to the distance from the beginning of the source sentence to current phrase position.

In fact, the bigger value of K is set1, the sparseness problem of data is more serious, especially for the maximum entropy classifier. In this paper, for illustrating simplicity without loss of generality, we set K = 1, therefore, the source word before the first phrase and the source word after the second phrase are adopted to be two additional features, i.e., S0.right and S3.left in Figure 3. Notice that if the first phrase is at the beginning of the source sentence, the S0.right feature will be set to "<s>", and if the second phrase is at the end of the source sentence, the S3.left feature will be set to "</s>". Compared to traditional eight features used in traditional methods, two additional features S0.right and S3.left are used by our method, as shown in Figure 3 with dotted frames. S0.right represents the right-most word of the phrases which is previous source phrase S1, and S3.left represents the left-most word of the phrases which is after the source phrase S2.

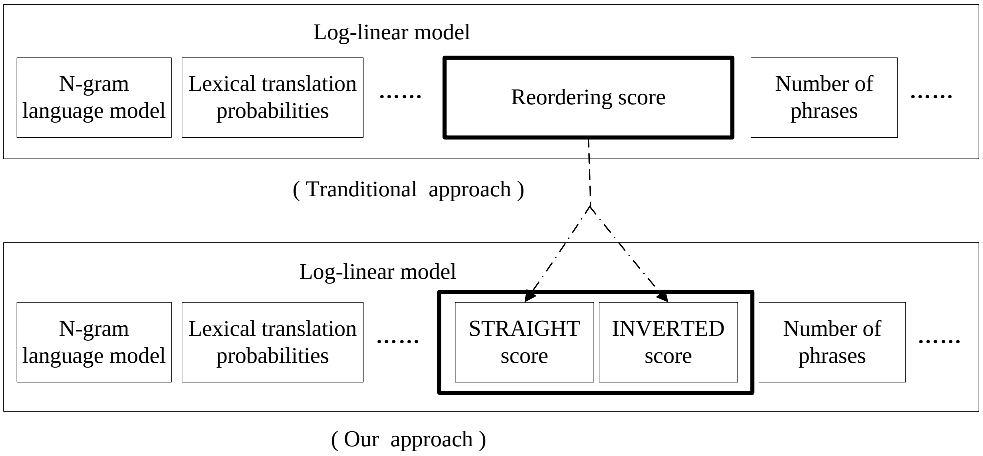


Figure 4. Reordering feature functions for decoding in two approachs

C. Reordering Feature Functions for Decoding

In statistic machine translation, all sub-models are trained separately and combined under the assumption that they are independent of each other in the log-linear model, the associated weightsλcan be tuned using minimum error rate training (MERT) (Och 2003). Base on the reordering samples and classification features, we can train a MaxEnt classifier to get the feature weights defined in section 3, and the reordering score is caculated by formular (3).

As we know, in traditional approach, the reordering is a sub-model which is in log-linear model, and the reordering score is used as a feature function. However, one feature function cannot indicate two phrase orientations. Therefore we define two feature functions to indicate two orientations. In this approach, we treat the reordering scores as two feature functions, STRAIGHT and INVERTED respectively.

The motivation behind this method is very simple: we want to depict the reordering model accurately in more dimensions to improve the discrimitive ability of the model. Taking the sentence mentioned in section 4.2 as an example, while the combination of two phrases which are "经济的" and "发展", the order of the consecutive phrases is predicted by the (maximum entropy) ME model to be STRAIGTH, then the reordering score is added to the STRAGTH score; The order of the combination "中国" and "经济 的 发展" is INVERTED, then the reordering score is added to the INVERTED score. The decoding algrithim repeats this operation to caculate STRAIGHT score and INVERTED score. After the whole source sentence is decoded, there are two reordering scores such as STRANGIT score and INVERTED score, they are integrated into the log-linear model and treated as two feature functions. The details can be illustrated by Figure 4.

V. EVALUATION

In this section, we compare the typical and our proposed methods within a phrase-based SMT system by experiments on the NIST Chinese to English translation tasks and Chinese to Japanses translation tasks.

A. Settings

The open source NiuTrans system (Xiao et al., 2012) was selected to build the baseline system. Our training corpus consists of 2 million sentences pairs in Chinese-English (Ch-En) task and 1.8 million sentences in Chinese-to-Japanese (Ch-Ja) task. Development data in Ch-En task is the NIST evaluation sets of mt04, and test data is the NIST evaluation sets of mt05 and mt06 2000 sentences are selected as development data and another 1200 sentences are selected as test data in Ch-Ja task. The base feature set used for all systems is similar to that used in (Marcu et al. 2006), including 14 base features in total such as 5-gram langusage model, bidirectional lexical and phrase-based translation probabilities. All features were combined log-linearly and their weights were estimated by performing minimum error rate training (MERT) (Och 2003).

B. Result

Observed from table 1, method 1, method 2 and method 3 show better performances than baseline with the increase of BLEU points on development set and test set in both Ch-En task and Ch-Ja task. Method 4 which integrates three methods synchronously shows significant improvements than baseline.

The improvements in method1 could be illustrated as follows. In traditional method, the reordering samples were learned base on word alignments, in other words, and it only consideres word alignment in current bilingual sentence pair, so it’s sensitive to the words alignment mistakes. Method 1 can alleviate this problem, it learns reordering samples from derivations of each bilingual sentence pair, and the derivations repesent the detais of how to generate the translation refference, therefore the quality of reordering samples is much higher than that of traditional method. In another view, this method considers the whole phrase table and chooses the phrase with the maximum model score when generating the translation hypothesis, so the word alignment mistakes in current sentence pair affect the training samples little in some extent. Comparing with baseline, method 2 considers both the information of the bilingual (source and target) phrases and the context of the two phrases in the source sentence, therefore the classifier could capture more contextual information and enhance the reordering prediction ability. Method 3 utlizes two feature functions to indicate the orientation during decoding, and show better performance than baseline.

TABLE I. IBM-BLEU4 (%) SCORE OF OUR METHOD ON DEVELOPMENT SET AND TWO TEST SETS ON TWO TASKS, * INDICATES SIGNIFICANTLY BETTER ON TEST PERFORMANCE AT THE P=0.05 LEVEL, COMPARE TO THE BASELINE METHOD.

		Ch-En		Ch-Ja	
Method	Description	Dev	Test	Dev	Test
Baseline	Baseline	39.83	33.27	30.11	25.40
Method1	Learning samples base on forced decoding	40.11 (+0.28)	33.58 (+0.31)	30.43 (+0.32)	25.67 (+0.27)
Method2	Boundary features and contextual features	39.94 (+0.11)	33.36 (+0.09)	30.24 (+0.13)	25.52 (+0.12)
Method3	STRAIGHT score and INVERTED score	40.03 (+0.2)	33.20 (+0.13)	30.40 (+0.29)	25.63 (+0.23)
Method4	using method1, method2 and method3 synchronously	40.34* (+0.53)	33.79* (+0.51)	30.74* (+0.63)	25.97* (+0.57)

TABLE II. THE COMPARATION ON THE NUMBER OF REORDERING SAMPLES EXTRACED BY TWO METHODS

Method	Number of STRAIGHT samples	Number of INVERTED samples	STRAIGHT / INVERTED
Base on word alignments (WA)	14.78 million	1.7 million	8.7 : 1
Base on forced decoding (FD)	10.58 million	2.46 million	4.3 : 1

In our approach, the reordering samples are extracted base on forced decoding, therefore the success rate of decoding influences the number of reordering samples. Table 2 lists the number of reordering samples by different method in Chinese to English bilingual sentence pairs (taking 1 million sentences as an example). In our experiments, when the beam size is set to 60, 24% of the bilingual sentence pairs fail to be decoded in the process of the forced decoding. In this case, for these failed sentence pairs, we adopt the results of traditional method for them. As shown in Table2, the number of samples base on WA is larger than that of FD, the ratio of STRAIGHT and INVERTED number reaches 8.7:1 and the ration of STRAIGHT and INVERTED numbers base on FD is 4.3:1, which is more preferable to the classifier, the distribution of the reordering samples is better than that of traditional method. The number of FD INVERTED reordering samples is larger than that of WA, the reason is that reordering samples are extracted from multiple derivation trees in a sentence pair.

To show the influence of our approach on translation compared with baseline, we present some examples which are listed in Table 3. Obviously, the translation result by our approach is better than that of baseline. In fact, the reordering model in this work influences the translation results which can be shown in two conditions. Firstly, during the decoding process, the reordering model in this work influences the selection of translation hypotheses and what we can see is the better translation result than baseline; Secondly, comparing with baseline, this model optimizes the phrases order in translation hypotheses and uses the same translation hypothesis with baseline, but better translation result shown for us.

VI. CONCLUSION

This paper presents three optimization techniques to improve classification-based reordering methods for PBMT, involving reordering sample generation for classifier training, feature selection for classification and reordering feature functions for decoding. To our best knowledge, we are the first to apply forced decoding technique to generate training samples on reordering and treat the reordering score as two feature functions into log-linear model. Experimental results show that the work in this paper improves the baseline system significantly. In future work, we can make research on extending the sliding window defined in this paper to capture more contextual information and utilize other models to improve the reordering for PBMT.

TABLE III. THREE EXAMLES OF TEST BY TRADITIONAL METHOD AND OUR APPROACH, PHRASE WHICH REPRESENTS THE VARIABLE POSITION IN DIFFERENT POSITION IS MARKED IN BOLD

Case 1	Chinese	同时 , 将 海军 新 装备 武器 试验 与 部队 科技 练兵 相结合 , 缩短 了 海军 新 装备 形成 战斗力 的 时间
	Baseline	Meanwhile, naval weapons testing of new equipment with the combination of the science and technology training of troops, to shorten the time new equipment to form combat the navy
	Our approach	Meanwhile, the combination of naval weapons testing of new equipment and the science and technology training of troops, to shorten the time new equipment to form combat the navy
Case 2	Chinese	组委会 和 国际 联盟 在 同一天 作出 了 对 她 禁赛 两年 的 处罚 决定
	Baseline	The organizing committee and the international union banned for two years a decision on the penalty made on the same day to her
	Our approach	The organizing committee and the international union banned for two years a decision on the penalty made to her on the same day
Case 3	Chinese	几天前，孩子模仿电视自杀了。
	Baseline	数日前、子供は模倣テレビが自殺です。
	Our approach	数日前、子供はテレビを模倣して自殺します。
Case 4	Chinese	水产厅资源管理部的负责人就该海域的情况进行了说明。
	Baseline	水産庁資源管理部の責任者が海域の状況を説明し、説明を行った。
	Our aproach	水産庁資源管理部の担当者は、当該海域の状況をこう説明する。

REFERENCES

[1] Li, P., Liu, Y., Sun, M., Izuha, T., & Zhang, D. (2014). A Neural Reordering Model for Phrase-based Translation. In COLING (pp. 1897-1907).

[2] Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2):263--311, June.

[3] Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human

[4] Language Technology - Volume 1, NAACL'03, pages 48--54, Stroudsburg, PA, USA. Association for Computational Linguistics.

[5] Koehn, P., Axelrod, A., Mayne, R. B., Callison-burch, C., Osborne, M., and Talbot, D. (2005). Edinburgh system description for the 2005 iwslt speech translation evaluation. In In Proc. International Workshop on Spoken Language Translation (IWSLT).

[6] C. Tillmann. 2004. A unigram orientation model for statistical machine translation. In Proceedings of HLT-NAACL: Short Papers, pages 101--104.

[7] Nagata, M., Saito, K., Yamamoto, K., & Ohashi, K. (2006, July). A clustered global phrase reordering model for statistical machine translation. InProceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the

Association for Computational Linguistics (pp. 713-720). Association for Computational Linguistics.

[8] D. Xiong, Q. Liu, and S. Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 521--528, Sydney.

[9] B. Xiang, N. Ge, and A. Ittycheriah. 2011. Improving reordering for statistical machine translation with smoothed priors and syntactic features. In Proceedings of SSST-5, FifthWorkshop on Syntax, emantics and Structure in Statistical Translation, pages 61--69, Portland, Oregon, USA. Association for Computational Linguistics.

[10] R. Zens and H. Ney. 2006. Discriminative reordering models for statistical machine translation. In Proceedings on theWorkshop on Statistical Machine Translation, pages 55--63, New York City, June. Association for Computational Linguistics.

[11] V. Nguyen, A. Shimazu, M. Nguyen, and T. Nguyen. 2009. Improving a lexicalized hierarchical reordering model using maximum entropy. In Proceedings of the Twelfth Machine Translation Summit (MT Summit XII). International Association for Machine Translation.

[12] Zens R, Ney H, Watanabe T, et al. Reordering constraints for phrase-based statistical machine translation[C]//Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004: 205.

[13] Alrajehab, Abdullah, and Mahesan Niranjanb. 2014.Bayesian Reordering Model with Feature Selection. ACL 2014, 477.

[14] Nieben, S., & Ney, H. (2001). Lehrstuhl Fur Informatik Vi. Morphosyntactic analysis for reordering in statistical machine translation.

[15] Collins, M., Koehn, P., Ku\v{c}erov\'{a}, I. (2005, June). Clause restructuring for statistical machine translation. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 531-540). Association for Computational Linguistics.

[16] Zang, S., Zhao, H., Wu, C., & Wang, R. (2015, August). A novel word reordering method for statistical machine translation. In Fuzzy Systems and Knowledge Discovery (FSKD), 2015 12th International Conference on (pp. 843--848). IEEE.

[17] Isozaki, H., Sudoh, K., Tsukada, H., & Duh, K. (2010, July). Head finalization: A simple reordering rule for sov languages. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR(pp. 244-251). Association for Computational Linguistics.

[18] Galley, M., & Manning, C. D. (2008, October). A simple and effective hierarchical phrase reordering model. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 848-856). Association for Computational Linguistics.

[19] Wu, D. (1996, June). A polynomial-time algorithm for statistical machine translation. In Proceedings of the 34th annual meeting on Association for Computational Linguistics (pp. 152--158). Association for Computational Linguistics.

[20] Och, F. J. 2003. Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), Sapporo, Japan, pp. 160-67

[21] Xiao T, Zhu J, Zhang H, et al. NiuTrans: An open source toolkit for phrase-based and syntax-based machine translation[C] Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012: 19-24

Optimal Pricing for Service Provision in IaaS Cloud Markets

Gang Fang
Trade Circulation Institute, Anhui Business College, Hefei, China, 230000;

Zhengce Cai
Department of Information Service,
Anhui Business College,
Hefei, China, 230000

Xianwei Li*
School of Information Engineering,
Suzhou University, Suhou, China, 234000;
*lixianwei163@163.com

Abstract —Pricing plays an important role for service provision in cloud computing. In this paper, we investigate price based resource access control in two Monopoly IaaS cloud market, respectively. The two IaaS cloud market is formed by one public cloud service providers (CSPs) and cloud broker (CB), provisioning cloud services to delay-sensitive cloud users (CUs). In the first monopoly cloud market, we treat the public CSP as an M/M/1 queueing system and study this CSP's pricing effect on the equilibrium behaviours of self-interested CUs. We propose two pricing mechanisms with the objective of maximizing revenue and social welfare, respectively. In the second monopoly cloud market, the CB is modelled as an M/M/∞ queueing system, which has infinite capacity to serve a common pool of CUs. We also analyze how pricing affects the equilibrium behaviors of CUs and the revenue-optimal and social-optimal pricing strategies in view of this CSP.

Keywords-Pricing, IaaS; cloud market; queueing system

I. INTRODUCTION

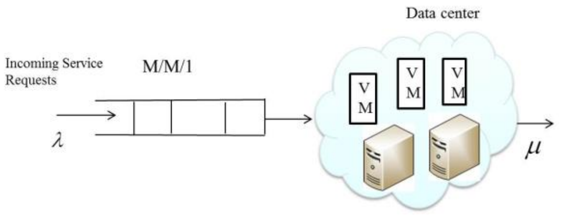
In recent years, cloud computing has received a significant amount of attentions from both engineering and academic fields and the use of cloud service is proliferating. Cloud computing can be defined by several ways, one widely adopted is proposed by Buyya et al. [1] :

"a cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and the consumers"

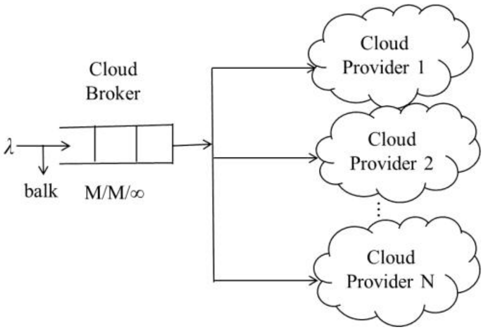
Cloud services are mainly classified into three types [2]: Infrastructure as a Service (IaaS), Software as a Service (SaaS) and Platform as a Service (PaaS). A recent study show that the market size of cloud computing will reach $112 billion in 2018, in a large part due to IaaS cloud services [3]. We focus on IaaS clouds in this paper, where CSPs deliver Infrastructure as a Service (IaaS) to cloud users. In the cloud computing environment, IaaS CSPs bundle their physical resources, such as CPU, memory and disk, into distinct types of virtual machine (VM) instances, according to their sizes and features, and offer them as services to users. Amazon EC2 is a public CSP which has hosted several types of VM instances (e.g. small, medium, large and extra large) based on the capacities of CPU, memory and disk [4], the configurations of some VM instances are shown in Table 1. Cloud users purchase units of computing time on VM instances to run their jobs.

Optimal pricing for cloud resources has been extensively studied by a significant amount of works in the literature. Feng et al. studied non-cooperative price competition in an oligopoly public cloud market [5]. They modelled each PCP as an M/M/1 queue, and analyzed how to set optimal prices in order to maximize the revenues of PCs based on resource capacities and the job finishing time. Xu et al., presented a study pricing cloud resources in a monopoly public cloud market [6]. Their study indicated that the revenue got with reserved pricing is not less than the first-order discrimination pricing. Mashayekhy et al. proposed a federation formulation game that considers the cooperation of these cloud providers to offer cloud services [7]. Their designed cloud federation mechanism enables cloud providers dynamically to form a federated cloud, which maximizes the profits of cloud providers.

In this paper, we study pricing-based service access control of CUs in a heterogeneous cloud market formed by two CSPs, public CSP and CB provisioning cloud services to delay-sensitive CUs. We consider two cloud scenarios corresponding to two types of cloud market: public CSP monopoly, and CB monopoly, which is illustrated in Figure 1. We note that similar structure analysis is also adopted by [9] in which the authors studied optimal pricing effects on the equilibrium behaviours of secondary users in cognitive radio networks. However, the effects of delay costs charged by CSPs on cloud users are not fully considered in [9]. By incorporating the delay costs of CSPs, in the first monopoly



(a) Public cloud monopoly market



(b) Cloud broker monopoly market

Figure 1. Two cloud market scenarios

Cloud market, we model the PC as an M/M/1 queueing system and analyze the pricing effect of this CSP on the equilibrium behaviours of non-cooperative delay-sensitive CUs. These behaviours are characterized by CUs' service access decisions of joining or balking to the queue upon arrival. From the viewpoint of CUs, their service access decision model are made according to the individual optimal strategy exploited by each CU, which is based on a utility function that captures the heterogeneous delay- sensitivity of CUs. We then show that there is a unique Nash equilibrium of CUs' joining probability in the non-cooperative game among them. In terms of the monopoly CSP, we design two pricing policies with the objective of maximizing revenue and social welfare, respectively.

In the second monopoly market, the CB is modelled as an M/M/∞ queueing system provisioning cloud services to delay-sensitive CUs. Similar to the first monopoly cloud market, we also study the CSP's pricing effect on the equilibrium behaviours of CUs. Since the CB has sufficient resources to serve the needs of CUs, therefore, it can provide better quality of service (QoS) measured by the average queueing delay. From the perspective of this CSP, we also study two pricing policies with the objective of maximizing revenue and social welfare, respectively.

The rest of the paper is structured as follows. System models are presented in section 2. We analyze the monopoly public cloud market in the section 3, the monopoly CB cloud market in the section 4. Conclusions and future works are given in section 5.

TABLE I. CONFIGURATIONS OF SOME AMAZON EC2 VM INSTANCES

Instance Types	**Compute Unit**	**Storage (GB)**	**Memory (GiB)**
c3.large	2	32SSD	3.75
c3.xlarge	4	80SSD	7.5
c3.2xlarge	8	160SSD	15
c3.4xlarge	16	32SSD	30
c3.8xlarge	32	80SSD	60

II. SYSTEM MODELS

A. CUs model

We assume that there is potential stream of CUs arrive at the cloud market with rate λ according to the Poisson process. Each CU carries a distinct job upon arrival. Therefore, we use CU and job interchangeably throughout the paper. The jobs of CUs in cloud data centers are classified into two types [10]: interactive (delay-sensitive) jobs, such as web service, and batch (delay-tolerant) jobs, such as scientific applications. Recent study shows that delay-sensitive interactive workloads take over 50% of data center workloads [11]. Hence, we focus on delay-sensitive interactive jobs and assume that each job attached to a specific application is denoted by a parameter θ, which reflects the sensitivity of CU's application to delay. The value of θ is private, but its distributions are known to CSPs. We also assume that θ is uniformly distributed on [0,1] with probability distribution function (PDF) f(.) and cumulative distribution function (CDF) F(.). This assumption is also widely adopted in the literature [9] [12][13].

When a type-θ CU arrives to the cloud market, it must make a decision as to whether to acquire service or not. If joins CSPi (i=p or c, where p and c denotes public CSP and CB, respectively), it will get net utility which is

$$U_i = r - \theta d_i - p_i, i = p, c \tag{1}$$

This net utility function is commonly used in the cloud and communication networks literatures [5][9][13], which captures the balance between the reward r and the total costs θcdi+pi that a CU takes if it joins the queueing system CSPi. The reward r represents the benefit factor of a CU for accessing the cloud service [9][13]. The total costs include two parts: θdi and pi, where di is the average queueing delay that this job experiences in the queueing system and pi is the price per service request charged by this CSPi. The similar pricing scheme is widely adopted by CSPs. Such as, Campaign Monitor [14] and Amazon Simple Email Service (ES) [15] charge CUs according to the number of campaigns and emails they process, respectively.

B. Public Cloud Service Provider (CSP)

When a type-θ CU decides to subscribe the service from the public provider, it will join a queueing system of this public provider. The system of the PC is modelled as an M/M/1 queue with service rate μ serving a potential number of CUs. The M/M/1 queue model is widely used in the cloud computing literature [9] to analyze response time as a function of the capacity of cloud resources and arrival rate of service requests. From (1), the net utility of type-θ CU for accessing the service public provider given price p1 is

$$U_1 = r - \theta c d_1 - p_1 \tag{2}$$

Where $d_1(\lambda) = 1/(\mu - \lambda)$ is the average queueing delay incurred by the arrival rate λ.

C. Cloud Broker (CB)

Since the CB integrates and coordinates resources among different CSPs, therefore, we assume that it has sufficient cloud resources to meet the demands of CUs. Hence, the system of CB is modelled as an M/M/∞ queue with enough servers to serve a common potential pool of CUs. The similar models have been widely used in the cloud literature to analyze power management or resource allocation in data centers. In [16], the authors studied optimal multi-server configuration to maximize profit of CSPs in cloud data centers. In [17], by modelling the CSP as M/G/m/m+r queueing system, the authors analyzed the performances of cloud data centers. Fang et al. studied throughput and energy tradeoff in mobile cloud platforms by applying the M/M/m queueing model [18]. From (1), the net utility of type-θ CU for accessing the service cloud broker given price p2 is

$$U_2 = r - \theta d_2 - p_2 \tag{3}$$

where $d_2 = 1/\mu$ captures the average queueing delay in M/M/∞ queue.

III. Public Cloud Monopoly Market

In this section, we first investigate the decisions of CUs as to whether to join or balk to the public provider and then design two optimal pricing mechanisms with the aim of maximizing revenue and social welfare, respectively.

A. CUs' Decision Policy

We consider a number of CUs arriving at the public cloud market, and these CUs are rational decision-makers in that they are only concerned with their own net utilities. Upon arrival, each type-θ CU has to make a decision whether to join or balk the queueing system of the public provider. It will join the queue if and only if its net utility U1(θ)≥0. Therefore, we get the following individual optimal decision policy.

Definition 1. A self-optimizing type-θ CU with its net utility $U_1(\theta) = r - \theta c d_1(\lambda_1) - p_1$ will follow a joining decision policy such that

- it joins public provider if U1(θ)≥0，which requires θ≤θ1, where

$$\theta_1 = \frac{r - p_1}{c d_1(\lambda_1)} \tag{4}$$

- it balks, if U1(θ)<0.

The above definition indicates that the fraction of CUs that have θ values less than θ1 will subscribe to the public provider. The fraction of CUs that have θ values less than θ1 is

$$F(\theta_1) = \int_0^{\infty} f(\theta) d\theta = \int_0^{\theta_1} f(\theta) d\theta \tag{5}$$

Then, the effective arrival rate of CUs to the public provider denoted by λ1 is

$$\lfloor 1 = \lfloor \Phi(\backslash 1) \tag{6}$$

B. Revenue Optimal Pricing Mechanism

Under the assumption that the public cloud provider knows the effective arrival rate, when charging p1 and delay cost c, this public cloud provider can get revenue $\pi_1(p_1) = p_1 \lambda_1$. The objective of the public cloud provider is to maximize its revenue, which can be formulated as

$$\mu\alpha\xi\, \pi_1(p_1) = p_1 \lambda_1 \tag{7}$$

$$\text{s.t. } p_1 \in [p_{\text{low}}, p_{\text{up}}]$$

where pup=r, plow=max{0,r-cd(λ1)}.

It is obvious that π1= p1λ1 is a concave function from $\pi_1''(p_1) < 0$. Hence, the problem of (7) can be solved by efficiently. By setting the first derivative $\frac{\partial \pi}{\partial p_1} = 0$, we get the optimal price

$$p_1^* = \frac{c + \lambda r - \sqrt{c(c + r\lambda)}}{\lambda} \tag{8}$$

Accordingly, the optimal revenue is

$$\pi_1^* = \lambda_1 p_1^* = \frac{\mu[2c + \lambda r - 2\sqrt{c(c + r\lambda)}]}{\lambda} \tag{9}$$

C. Social Welfare Optimal Pricing Mechanism

Cloud social welfare is the net utilities of CUs plus the revenue of the public cloud provider. When charging price p1 , only the fraction of CUs with θ≤θ1 subscribe to the public loud provider. Therefore, the cloud social welfare at price p1 is

$$\begin{aligned} S_1(p_1) &= U_1 + \pi_1 \\ &= \int_0^{\theta_1} [r - \theta c d_1(\lambda_1)] f(\theta) d\theta \\ &= r\theta_1 - \frac{\theta_1^2 c}{2(\mu - \lambda\theta_1)} \end{aligned} \tag{10}$$

where θ1 is given in (4). From (4) we know that θ is the function of price p1. Therefore, the variable of $S_1(\theta_1)$ can be changed from p1 to critical CU variable. Hence, the social welfare optimal pricing problem is formulated as

$$\mu\alpha\xi\, S_1(\theta_1) \tag{11}$$

s.t. $\theta_1 \in [0, 1]$

where $S_1(\theta_1)$ is given in (10).

We find that the objective function of problem (11) is concave by calculating $S_1''(\theta_1) < 0$, therefore, the optimal solution of (11) can be effectively solved, which is denoted by θ_1^S. Hence, the optimal social welfare price is

$$p_1^S = r - \theta_1^S c d_1(\lambda \theta_1^S) \tag{12}$$

IV. Cloud Broker Monopoly Market

In this section, we first investigate the decisions of CUs as to whether to join or balk to the cloud broker and then design two optimal pricing mechanisms with the goal of maximizing revenue and social welfare, respectively.

A. *CUs' Decision Policy and Equilibrium*

We consider a number of CUs arriving at the federated cloud market, and these CUs are rational decision-makers in

f CUs that have θ values less than θ2 will subscribe to the cloud broker. The fraction of CUs that have θ values less than θ2 is expressed as

$$F(\theta_2) = \int_0^{\infty} f(\theta)d\theta = \int_0^{\theta_2} f(\theta)d\theta \tag{14}$$

Then, the effective arrival rate of CUs to the public provider denoted by λ2 is

⌊2=⌊Φ(⌊2) (15)

B. *Revenue Optimal Pricing Mechanism*

Under the assumption that the cloud broker knows the actual arrival rate of CUs, when charging p2 and delay cost c, this CSP can get revenue $\pi_2(p_2) = p_2\lambda_2$. The objective of the cloud broker is to maximize its revenue, which can be formulated as

μαξ $\pi_2(p_2) = p_2\lambda_2$ (16)

s.t. $p_2 \in [0, r]$

By setting the first derivative of the objective function with respect to p2 to zero, we get the revenue optimal price

$$p_2^* = \frac{r}{2} \tag{17}$$

Accordingly, the optimal revenue is

$$\pi_2^* = \frac{\mu r^2 \lambda}{4c} \tag{18}$$

that they are only concerned with their own net utilities. Upon arrival, each type-θ CU has to make a decision whether to join or balk the queueing system of the cloud broker. For a CU, it will join the queue if and only if its net utility U2(θ)≥0. Therefore, we get the following individual optimal decision policy.

Definition 2. A self-optimizing type-θ CU with its net utility $U_2(\theta) = r - \theta c d_2(\lambda_2) - p_2$ will follow a joining decision policy such that

- it joins public provider if U2(θ)≥0，which requires θ≤θ2, where

$$\theta_2 = \frac{r - p_2}{c d_2} \tag{13}$$

- it balks, if U2(θ)<0.

The above definition indicates that the fraction o

C. *Social Optimal Pricing Mechanism*

The cloud social welfare is defined as

$$\begin{aligned} S_2(p_2) &= U_2 + \pi_2 \\ &= \int_0^{\theta_2} [r - \theta c d_2] f(\theta) d\theta \\ &= r\theta_2(p_2) - \frac{c(\theta_2(p_2))^2}{2\mu} \end{aligned} \tag{19}$$

The cloud social welfare problem is formulated as

μαξ $S_2(p_2)$ (20)

s.t. $p_2 \in [0, r]$

By setting the first derivative of the objective function with respective to p2, the socially optimal price is

$$p_2^S = r - c\mu \tag{21}$$

Acknowledgment

This paper is supported by the following projects, Anhui Key research projects of Humanities and Social Sciences (SK2016A0207), and Suzhou Regional Collaborative Innovation Center (2016szxt05).

References

[1] R. Buyya, C.S. Yeo, and S. Venugopal, "Market Oriented Cloud Computing: Vision, Hype, and Reality for Delivering it Services as Computing Utilities," Proc. 10th IEEE Conference on High Performance Computing and Communications (HPCC 2008), pp. 5-13, Sept. 2008.

[2] D. Bruneo, “A stochastic model to investigate data center performance and QoS in IaaS cloud computing systems,” IEEE Trans. Parallel Distrib. Syst., vol.25, no.3, pp.560–569, March 2014.

[3] L. Zheng, Carlee Joe-Wong, and C. G. Brinton et al. “On the Viability of a Cloud Virtual Service Provider,” Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science (SIGMETRICS 2016), Antibes Juan-les-Pins, France, pp. 235-248, June 2016.

[4] Amazon EC2 Pricing. http://aws.amazon.com/cn/ec2/pricing/.

[5] Y. Feng, B. Li, and B. Li, “Price competition in an oligopoly market with multiple IaaS cloud providers,” IEEE Trans. Comput., vol. 63, no. 1, pp. 59-73, Jan. 2014.

[6] H. Xu and B. Li, “A study of pricing for cloud resources,” ACM SIGMETRICS Performance Evaluation Review, vol. 40, no. 4, pp. 3-12, Mar. 2013.

[7] L. Mashayekhy, M. M. Nejad, and D. Grosu. "Cloud federations in the sky: Formation game and mechanism." IEEE Trans. Cloud Comput., vol. 3, no. 1, pp.14-27, Jan.-March 2015.

[8] Z. Liu, Y. Chen, and C. Bash, et al., “Renewable and cooling aware workload management for sustainable data centers,” Proc. of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems (SIGMETRICS 2012), London, England, UK, pp. 11–15, June 2012.

[9] N. H. Tran, C. S. Hong, and S. Lee et al., “Optimal Pricing Effect on Equilibrium Behaviors of Delay-Sensitive Users in Cognitive Radio Networks,” IEEE J. Sel. Areas Commun., vol. 31, no. 11, pp. 2266–2579, Oct. 2013.

[10] Z. Liu, M. Lin, and A. Wierman, et al., “Greening geographical load balancing,” IEEE/ACM Trans. Netw., vol. 23, no. 2, pp. 657-671, Apr. 2015.

[11] Y. Jin, S. Sen, and R. Guerin, et al., “Dynamics of competition between incumbent and emerging network technologies,” in Proc. Workshop on the Economics of Networks, Systems, and Computation (NetEcon’ 08), Seattle, WA, USA, pp. 49–54, Aug. 2008.

[12] R. Gibbens, R. Mason, and R. Steinberg, “Internet service classes under competition,” IEEE J. Sel. Areas Commun., vol. 18, no. 12, pp. 2490–2498, Dec. 2000.

[13] Campaign Monitor. http://www. campaignmonitor.com/pricing

[14] Amazon SES. https://aws.amazon.com/ses /pricing.

[15] J. Cao, K. Hwang, and K. Li, et al., “Optimal multiserver configuration for profit maximization in cloud computing,” IEEE Trans. Parallel Distrib. Syst., vol.24, no.6, pp.1087-1096, June 2013.

[16] H. Khazaei, J. Misic, and V. B. Misic, “Performance analysis of cloud computing centers using M/G/m/m+r queuing systems,” IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 5, pp. 936–943, May 2012.

[17] W. Fang, Y. Li, and H. Zhang, et al. "On the throughput-energy tradeoff for data transmission between cloud and mobile devices." Information Sciences, 283, pp.79-93, 2014.

[18] Fudenberg and J. Tirole, Game Theory, MIT Press, Cambridge, USA, 1991.

Lidar Image Classification based on Convolutional Neural Networks

Yang Wenhui*,
School of computer science and engineering Xi'an
Techonolgical University,Xi'an ,Shananxi 710021
e-mail:1353678463@qq.com

Yu Fan
School of computer science and engineering Xi'an
Techonolgical University,Xi'an ,Shananxi 710021
e-mail:yffshun@163.com
*The correspoding author

Abstract—This paper presents a new method of recognition of lidar cloud point images based on convolutional neural network. This experiment uses 3D CAD ModelNet, and generates 3D point cloud data by simulating the scanning process of lidar. The data is divided into cells, and the distance is represented by gray values. Finally, the data is stored as grayscale images. Changing the number of cells dividing point cloud results in different experimental results. Experiments show that the proposed method has higher accuracy when dividing the cloud with 27×35 cells. Comparison of point cloud cell image method with VoxNet method, *experimental results show that the classification method based on gray image and convolutional neural network has more advantages than the most advanced point cloud recognition network Voxnet.*

Keywords-Point Cloud; CNN; Gray Image; Lidar;

I. INTRODUCTION

Lidar has been widely used in the acquisition of point cloud data because of its high precision and wide range of visibility. The classification and recognition of point cloud images formed by lidar has been the focus of many domestic and foreign famous experts and scholars. The key technology and the final aim of this method are feature extraction and classification of point cloud images.

Domestic and foreign famous experts and scholars have done a lot of research work on it. Some scholars use manual extraction of features, and then select a classifier for classification and recognition methods. R. B. Rusu and others use the relation between the normal vectors of a region as feature [1], and classify objects by classification. Yasir, Salih and others use VFH as a feature, and SVM is used as classifier to classify and identify point clouds. [2]. Liu Zhiqing and others improved classifier, and used information vector machine to classify and identify point cloud [3]. The laser scanning point cloud data, sparse and incomplete, effective and accurate description of artificial feature selection is often difficult. n Researchers need professional knowledge and heuristic methods, which rely on personal experience to a great extent. While deep learning can automatically extract features and classify them, they are invariant to displacement, scaling, and other forms of rigid body change. Therefore, in recent years, some experts and scholars have begun to use deep learning to classify point cloud images. Daniel, Maturana and Sebastian Scherer processed the data into a physical form, the feature was extracted by using the three-dimensional convolution kernel, and the neural network was used to classify and recognize [5]. Good results were obtained.

Lidar data is difficult to obtain relative to visible image data. Some experts use Sydeny Urban Objects data sets to train the network, but this dataset is too small. It can not achieve the purpose of data-driven. Therefore, this experiment uses the ModelNet data sets with rich data types as the basis, and preprocess the data, then use the convolution neural network to classify and recognize. The VoxNet recognition method proposed by Daniel, Maturana and others achieves higher accuracy in the 3D data set, while the cloud point image is not fully used in the recognition of the laser point cloud image, thefore the accuracy is not very high.

In order to compare, the ModelNet data set is classified and identified by using the VoxNet method and the point cloud cell image proposed in this experiment. By experiment , the cloud cell image method achieves higher accuracy.

II. MODEL CONSTRUCTION BASED ON CONVOLUTIONAL NEURAL NETWORK

In this paper, the formation and principle of convolutional neural networks are briefly introduced. The network structure of VoxNet and point cloud cell images is compared. The experimental scheme is designed, and the accuracy of the two networks is compared. Some useful conclusions are obtained.

A. Convolutional neural network

Convolutional neural network is inspired by the neural mechanism of visual system, it is a kind of deep learning ability of artificial neural network system. Compared with the traditional method, convolution neural network has the advantages of strong applicability, feature extraction and classification at the same time, strong generalization ability, can be used to recognize the change of rigid body displacement, zoom and other forms about two-dimensional or three-dimensional image. It has become the focus of the field of machine learning at present [6].The standard convolutional neural network is a special multilayer feedforward neural network. It has a deep network structure, which is generally composed of input layer, convolutional layer, down sample layer, full connection layer and output layer. Among them, the Convolutional layer, the down sample layer and the full connection layer are hidden layers. In the convolutional neural network, input layer is usually a matrix for receiving original image; convolution layer for image feature extraction, reduce the noise interference; convolution layer sharing local weights, the special structure is more close to the real biological networks, the CNN has a unique advantage in the field of image processing. Compared with the full connection layer, the shared weight reduces the network parameters and accelerates the training speed. On the other hand, the complexity of the network is reduced, and the multidimensional input signals (voice and image) can be input directly, thus avoiding the process of data rearrangement during feature extraction and classification [6]. The down sample layer reduces the amount of data to be processed according to the principle of local correlation of the image, and the output layer maps the extracted features to the predicted tags.

Convolution of convolutional maps in convolution neural networks is discrete and can be expressed as Eq.（1）：

$$x_{\beta}^{\gamma} = f\left(\sum\nolimits_{\alpha\epsilon M_{\beta}} x_{\alpha}^{\gamma-1} k_{\alpha\beta}^{\gamma} + b_{\beta}^{\gamma}\right) \qquad （1）$$

The weight and bias of CNN can be learned by back propagation algorithm, so it is not necessary to extract features manually. The convolution neural network uses the classical BP (back propagation) algorithm to adjust the parameters, and finally completes the learning task. BP network update weights for Eq.（2）:

$$\omega(n+1) = \omega(n) - \eta \frac{\partial C}{\partial \omega} \qquad （2）$$

$\omega(n)$ represents the nth map , and $\omega(n+1)$ represents the (n+1)th map. η represents Learning rate. $\frac{\partial C}{\partial \omega}$ represents the loss function, can be obtained by back propagation.

B. Model comparison

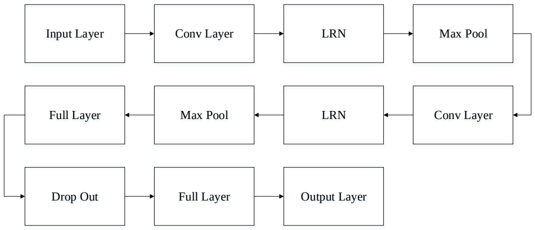


Figure 1. Point cloud cell network structure diagram

Figure. 1 is a convolutional neural network structure for testing in this paper. A total of two volumes are formed. The number of feature maps is 10 and 15, and the size of the convolution kernel is 8 and 5.Each convolutional layer has a 2 × 2 maximum pool layer for preventing over fitting and a LRN layer for local normalization. The discard rate of dropout layer is 0.5. The number of neurons in two fully connected layers was 256 and 10, respectively.

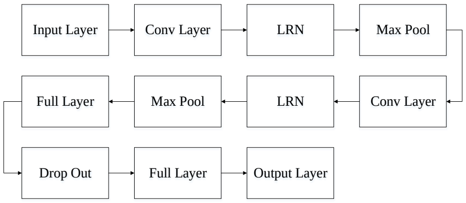


Figure 2. VoxNet network structure diagram

Figure. 2 is a network structure of the VoxNet, and the input layer accepts data in the form of $32 \times 32 \times 32$. A total of two volumes are laid, and the number of feature maps is 32, using $5 \times 5 \times 5$ and $3 \times 3 \times 3$ convolution kernels, respectively. The discard rate of the Dropout layer is 0.2 and 0.3, which can prevent overfitting and reduce the amount of computation. The largest pool layer uses $2 \times 2 \times 2$ filter. Finally, there is a full layer of neurons with a number of 128 and a dropout layer with a discard rate of 0.4. The seventh layer is the output layer, and the number of neurons is 10.

VoxNet divides the processed data into $32 \times 32 \times 32$ cells, which will lose some of information. In this paper, the data is projected into a point cloud cell image, which can save all the data and improve the utilization of data, thus improving the classification accuracy.

III. EXPERIMENTS

A. Experiment environement and Datasets

We use TensorFlow-gpu 0.12.1 open source software library, Windows 7 operating system, NVIDIA GTX 950 graphics card in our experiments. The experiment uses data for Princeton University's ModelNet which is a large 3D CAD model database similar to the ImageNet. ModleNet10 is a subclass of ModelNet that contains 3991 CAD models and 10 categories, all of which are distinct and located in the middle. This paper uses ModelNet10 as the experimental data set.

B. Reference frame and resolution

ModelNet stores data as OFF format, and OFF format represents the geometric structure of a model by describing polygons on the object surface. We save OFF format data as STL format by 3ds Max, STL using triangle to represent an object model is more conducive to the data processing in this experiment. We simulate the scanning process of laser radar, and use the gravity center method to determine whether the laser is scanned on the triangle [11], thus get a three-dimensional image with only one side. VoxNet divides the data into $32 \times 32 \times 32$ cells. The proposed method in this study need different treatment to the data, firstly using the plane grid of $216 \times 280, 108 \times 140, 54 \times 70, 27 \times 35$ to partition the point cloud data, the formation of the matrix corresponding to the size, because each grid unit may correspond to multiple point cloud data, each matrix element values for the corresponding grid unit is all the average distance of point cloud. In order to compare the recognition effect under different partition methods, the matrix of different partition method is transformed into a gray image with resolution of$216 \times 280, 108 \times 140, 54 \times 70, 27 \times 35$ and the point cloud cell image is obtained, as shown in Figure 3.

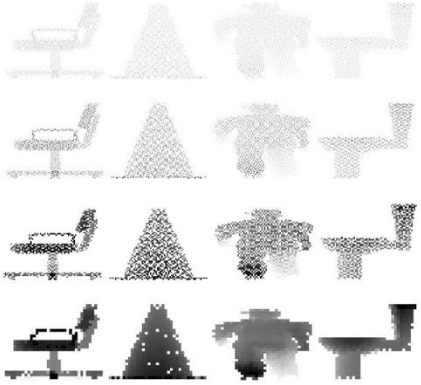


Figure 3. Gray image of point cloud divided by different cells Recognition result

When the number of cell division, and the point cloud image sparse number of point cloud, it is easy to produce some grid corresponding to zero, which represents the corresponding grid distance value is zero, but the distance of adjacent grid corresponding to the actual situation value should be smooth distribution. If used directly in network training, there will be greater deviations. In order to get better results, the experimental data will be divided into different cells, were divided into 216×280 , $108 \times 140, 54 \times 70, 27 \times 35$ cells in different partition methods have different results. The accuracy of the experiment, as shown in Table 1, can be seen:

TABLE I. DIVIDES THE ACCURACY RATE INTO DIFFERENT CELLS

Number of cells	accuracy rate
216 × 280	0.735
108 × 140	0.808
54 × 70	0.776
27 × 35	0.858

1) *The cells are divided differently, and the results are different. When the image is mapped to*216 × 280 *cells, the accuracy is lowest, and the accuracy is the highest when it is mapped to* 27 × 35 *cells. As the cells are smaller, the number of pixels in each cell increases and the number of points in the space is less, so a higher accuracy can be achieved.*

2) *The* 108 × 140 *cells have higher accuracy than* 54 × 70 *cells, because the distribution of the points in the* 108 × 140 *cells is more uniform, and better results can be obtained.*

When a point cloud is divided into grayscale images of different cells, its accuracy rate with the number of training times as shown in Figure 4. At the beginning of the network training, the network model is not fully trained, so the accuracy is low. With the increase of the number of iterations, the parameters of the network are also learning constantly, and the accuracy of classification recognition will gradually increase. Finally, the classification accuracy fluctuates in a small range, which means that the convolutional network is convergent and the classification accuracy is stable.

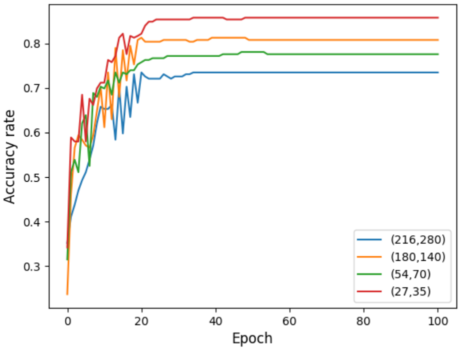


Figure 4. Relation between training times and recognition rate

C. *Comparison of point cloud cell image method with VoxNet method*

In order to test the accuracy of the experimental method, compare with VoxNet network which performs well in 3D recognition. VoxNet uses 3D point cloud data to classify and convert them into stereo cells, and the proposed network uses 3D point cloud data to divide gray images into input data. The accuracy of VoxNet classification recognition is 78.5%, and the recognition rate of the best segmentation results of cloud cell images is about 6% higher than the VoxNet. The experimental results are shown in table 2.

TABLE II. COMPARISON OF ACCURACY BETWEEN POINT CLOUD CELL IMAGE AND VOXNET RECOGNITION

Network type	VoxNet	Point cloud cell image
Recognition accuracy	0.785	0.858

IV. CONCLUSION

A classification and recognition method of point cloud based on convolutional neural network is proposed in this paper. Firstly, the point cloud data is processed as point cloud cell image, and then the network is used to classify and recognize the image. Experiments show that the classification method based on gray image and convolutional neural network has more advantages than the most advanced point cloud recognition network Voxnet.

REFERENCES

[1] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In Robotics and Automation, 2009. ICRA'09. IEEE International Conferenceon, pages 3212–3217. IEEE, 2009.

[2] Yasir Salih, A.S. Malik, D. Sidibé, M.T. Simsim, N. Saad and F. Meriaudeau .ompressed VFH descriptor for 3D object classification. 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2014.

[3] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015, pp. 1912–1920.

[4] Daniel Maturana,Sebastian Scherer.VoxNet: A 3D Convolutional Neural Network for real-time object recognition. Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on.

[5] J. Bergstra, F. Bastien, O. Breuleux, P. Lamblin, R. Pascanu, O. Delalleau, G. Desjardins, D. Warde-Farley, I. Goodfellow, A. Bergeronet al., "Theano: Deep learning on gpus with python," in NIPS 2011,BigLearning Workshop, Granada, Spain, 2011.

[6] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang and J. Xiao 3D ShapeNets: A Deep Representation for Volumetric Shape Modeling Proceedings of 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015).

[7] The HDF Group. Why Use HDF?. Retrieved January 4, 2012, from https://www.hdfgroup.org/why-hdf/.

[8] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. Guibas.Volumetric and multi-view cnns for object classification on 3d data. In Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2016.

[9] http://www.cnblogs.com/graphics/archive/2010/08/05/1793393.html The Princeton ModelNet. http://modelnet.cs.

[10] The Princeton ModelNet. http://modelnet.cs.

[11] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proc. CVPR, 2014.

Design and Development of Intelligent Logistics System Based on Semantic Web and Data Mining Technology

Yi Wang*
Department of Information Engineer
Guangdong Polytechnic
Foshan, China
E-mail: wangyifsfz@163.com

Haoyuan Ou
Department of Information Engineer
Guangdong Polytechnic
Foshan, China

Xue Bai
Department of Economic management
Guangdong Polytechnic
Foshan, China

***Abstract*—The intelligent logistics distribution of e-commerce is the computer technology and modern hardware equipment, software system and advanced management tools used by the logistics distribution enterprise. Data mining technology is the process of finding the probability distribution of random variables from a large number of source data. Automation of intelligent logistics system can improve labor productivity and reduce the error of logistics operation. This paper proposes design and development of the electronic commerce intelligent logistics system based on combination of semantic web and data mining. Experiments show that the proposed method is effective.**

Keywords-Semantic network; Intelligent logistics system; Electronic commerce; Data mining; Ontology

I. INTRODUCTION

Data mining should be applied to any kind of knowledge mining in any information storage mode, but the challenge and technology will be different because of the different types of source data [1]. In particular, the research shows that the data storage types of data mining is more and more abundant, in addition to some common value of the model, architecture and other research, but also carried out some research on the complex or new data storage methods of mining technology or algorithm.

The semantic web is a vision of the future, in which information is given a clear meaning that machines can automatically process and integrate information on the web. The Semantic Web uses XML to define custom tag formats and the flexibility to express data. Resource description framework (RDF) is a standard for describing resources. What is resource? This is a very difficult question to answer.

Ontology is used to describe the knowledge of a domain. The ontology describes the relations between concepts and concepts in the field. Different ontology languages provide different features [2]. The latest ontology language is the introduction of OWL W3C. It has plenty of operators, such as and, or, and negation. It is based on the logic that allows it to define concepts or concepts. Complex concepts can be built on simple concepts.

E-commerce logistics, also known as online logistics, is based on the Internet technology, creative aims to promote the development of the logistics industry, new business models, through the Internet, logistics companies to owner customers is a greater range of take the initiative to find, can in the country and even the world to expand the business, trading companies and factories can more quickly find the price the most suitable logistics company; e-logistics is committed to the world within the scope of the maximum number of a logistics needs of the owner of the business and provide logistics service of the logistics company are drawn together and help logistics supply and demand sides, deal with neutrality and integrity, free online logistics market. At present, there are more and more customers through the online logistics market to find a customer, to find a partner, to find a foreign agent. Online logistics provides the greatest value, is more opportunities. In this paper, we propose the design of intelligent logistics system based on Semantic Web technology and data mining.

II. ANALYSIS OF RDF TECHNOLOGY BASED ON SEMANTIC WEB

URI is the uniform resource identifier, which is used to identify and locate resources on the network. URI has many forms and can be extended, the most common of which we are familiar with the URL refers to the Internet is currently the best search engine Google refers to the founder of the semantic web, Berners Tim lee. In addition, there are many other forms of URI, including UUID, TAG and els, etc.. We can use URI to uniquely identify any thing, and any of the things that have URI can be said to be in the Web. For example, you just bought the book last week, some of your mind in the immature mind and even your own, etc., you can use URI on the network to identify.

Semantic web is not only able to understand human language, but also can make the communication between human and computer become as easy as the exchange between people. Adds more to the vocabulary of attributes and types, such as (cardinality), disjointness, equivalence, attributes, attributes (e.g., symmetry), and enumerated (classes).

Definition 1 [3]: the measure word limit is made up of three parts, respectively (quantifier), attribute and filler. There are someValueForm (some, least one at) and allValueFrom (OWL). Note that the universal measure word also describes those individuals who do not have the attribute, and it is the full name does not indicate the existence of a relationship, as is shown by equation (1), but that it must be associated with a particular class.

$$MI_{FA}(f,a)=\sum_{f,a}P_{FA}(f,a)\ln\frac{P_{FA}(f,a)}{P_F(f)P_A(a)} \tag{1}$$

Some more complex queries, such as a lot of filtering steps, are more important than the. But the traditional method is not competent for this. In addition, the Jena system attempts to create a list of features from the RDF data, which will be a collection of information on a number of features on a column topic, but it is not good for those who cannot come from a table.

Schema XML is actually a kind of XML application; it uses XML syntax itself, so the XML document is a kind of self description document. Schema XML is an alternative to DTD (Type Definition Document), but it is more flexible than DTD. It not only provides a complete set of mechanisms to restrict the use of tags in XML documents, but also supports more data types, and can provide a better data verification mechanism for the effective XML document service.

We find that DAML+OIL can easily express domain specific knowledge related to software requirements document, but not so suitable for the generation of domain models. We have investigated the use of DAML+OIL to describe non functional requirements such as quality of service, which is believed to be a meaningful training. We will continue to observe the development of language in semantic web, looking for the opportunity to combine generative rules and explore two levels of grammar as another possible language of the semantic web. Of course, as is shown by equation (2). TLG is able to obtain the semantic information associated with the combination of software components, but it can not be obtained by using the semantic web language [4].

$$fresp(x,y)=Det(Z)-kTrace^2(Z) \tag{2}$$

Definition 2: RDF is a powerful place, it only provides the main - that - the - object - the description of the form, as the predicate and object in the end is what, can be freely chosen according to different needs. Therefore, RDF can be defined as "resource description framework” and not "resource description method"". The most common predicate and object for RDF.

$$\begin{aligned}S_x&=E[(Y-EY)(Y-EY)^T]\\&=E[AX-E(AX)][AX-E(AX)]^T\\&=E[(A-EA)X][(A-EA)X]^T\end{aligned} \tag{3}$$

Similarly, we are now witnessing the early stages of XML's popularity. Although the XML itself is not sufficient for the realization of the semantic web, it is an important first step. XML and RDF are the W3C standards which are related to the semantic web. The earliest users may be interested in knowledge management and business to business electronic commerce. This momentum will drive more and more tools vendors and end users to adopt this technology. Of course, semantic web can be expressed as factual knowledge, and it can also be expressed as the connection between factual knowledge. Mainly in the following several aspects: 1, to express the fact; 2, to express the relation between things; 3, said the more complex knowledge.

Definition 3: RDF/XML how to put a RDF figure encoding XML elements attributes element content and attributes values of the basic ideas. URIrefs is written by XMLQNames, which consists of a short prefix (prefix) (representing a namespace URI) and an internal name (name local) (representing the elements or attributes of a namespace).

Semantic network is used to express complex concepts and their relations, so as to form a semantic network, which is composed of nodes and arcs. From the point of view of graph theory, they are a "directed graph", which is composed of nodes and nodes.

III. Research on Construction of Intelligent Logistics System of Electronic Commerce

The basis of network in logistics field is information, which means that the network has two meanings: first, the computer communication network, which includes the logistics distribution center and the supplier or manufacturer [5]. Two is the organization's network, namely so-called enterprise internal network (Intranet). Logistics network is the inevitable trend of logistics information, is one of the main characteristics of the logistics activities of e-commerce. In today's world, the availability of Internet and other global network resources and the popularization of network technology provide a good external environment for the logistics network, logistics network can not stop.

According to the information, we analyze the logistics distribution center at home and abroad. The conclusion is that they are all over the stage of simple delivery, which is essentially a real logistics distribution, but at the level of logistics distribution, which is in the primary stage of logistics distribution, it is not equipped with information, modernization and socialization. It is gratifying that the relevant departments of the country have recognized these problems, is from the macro level to guide the efforts of China's logistics and distribution industry in the information, modernization, socialization of the new logistics distribution direction, some of the government officials, business circles, academic circles are also common in this area, and have begun to practice.

Definition 3:Logistics center management information automation: in the construction of logistics center, we should make full use of modern information technology, such as bar code technology, radio frequency identification (RFID) technology, EDI (electronic data exchange technology), EOS (electronic ordering system), POS (point of sale), and so on, real information is automatic, fast and accurate collection, storage, transmission, processing and processing, logistics center for the management and intelligent monitoring to provide timely and reliable information support, as is shown by equation(4).

$$\Phi(m):=\left[\left(\prod_{i=0}^{0}A(mM+i)\right)^{T},\left(\prod_{i=1}^{0}A(mM+i)\right)^{T},\cdots,\left(\prod_{i=M-1}^{0}A(mM+i)\right)^{T}\right]^{T} \quad (4)$$

The development of electronic commerce and enterprise distribution system is closely related. A complete electronic commerce is a business activity which is composed of inventory, logistics, capital, payment, etc.. Logistics is an important part of the operation of electronic commerce [6]. It plays an important role in the electronic commerce. With the rapid development of electronic commerce, the requirements for logistics distribution are also higher and higher. In China, although there are a number of postal courier and some express company, but they only completed the most basic logistics system, and the time is long, the channel is not smooth, in the customer reputation is not good, simply can not meet the requirements of high efficiency, low cost of e-commerce logistics, seriously restricting the rapid development of China's e-commerce.

First, the application of automatic positioning and tracking of the vehicle. Dynamic information using the computer management of GPS system stops can through GPS and computer network real-time collection of railway car goods, vehicles, cargo tracking management, and vehicle scheduling management in a timely manner. Second, the management of railway transport. Using GPS of the computer information management system, through GPS and computer network real-time collection of railway trains, and it is locomotives, vehicles, containers and the goods of dynamic information, to realize the trains and cargo tracking management. As long as know truck vehicles, models and license plate number, you can immediately from nearly 10 million km of railway network flow of hundreds of thousands of trucks found in the truck, but also know the truck now where to run or stop in where, and all of the vehicle for the delivery of information.

Sorting automation: for the high demand, sorting operation of large quantities of goods to adopt mature automatic sorting technology, and it is so that can improve the efficiency of sorting, and can reduce labor intensity. In the use of automatic sorting, must be manual sorting to stay in operation space, in order to use a flexible manual sorting as a supplement to the peak demand. Packaging automation: the use of highly efficient and environmentally friendly packaging system, both to improve the packaging efficiency, but also to effectively reduce the volume of packaging. Such as Savoye's JIVARO is automatic packaging line.

E-commerce is helpful to improve the level of logistics management, electronic commerce, logistics information, logistics information, information processing, electronic, information transmission standardization and real-time, digital information storage, etc. [7]. With the development of logistics information technology, bar code technology, database technology, electronic ordering system and other technology, is gradually in the logistics field to get a wide range of applications, which will improve the level of logistics management to a certain extent, as is shown by equation (5).

$$V_j = V_{j-1} \oplus W_{j-1}, \quad \forall j \in Z \quad (5)$$

The traditional logistics distribution of the link is due to many subjects and the relationship of the artificial, so extremely cumbersome. In the e-commerce logistics distribution mode, the logistics distribution center can make the process simple and intelligent through the network. For example, the computer system management can make the whole logistics distribution management process becomes simple and easy to operate; the business promotion on the network platform can be used to make the shopping and trading process more efficient and less cost; logistics information is easy and effective dissemination of information makes the user to find and speed up the speed of decision-making process. Many of the activities that need to be processed and spent more time in the past are simplified because of the simplification of the network system, which greatly improves the efficiency of logistics distribution.

The automation of logistics system can improve labor productivity and reduce the error of logistics operation, but also can facilitate the collection and tracking of logistics information, improve the management and monitoring of the whole logistics system [8]. The facilities are very many, such as bar code automatic identification system, automatic sorting system, automatic access system, automatic guided vehicle system, cargo tracking system, etc.. In our country, the distribution of the logistics automation equipment should be determined according to the operating conditions and the characteristics of the goods, and the input and output analysis should be carried out in detail, as is shown by equation (6).

$$w(n+1) = w(n) + \frac{1}{2}\mu[-\nabla(E\{\varepsilon^2(n)\})] \quad (6)$$

Based on Semantic Web technology and data mining design of electronic commerce intelligent logistics system to complete the warehouse management of a single document management, including the storage of single edit and audit operation, the system must be able to maintain the data table of these operations. At the same time to maintain the information contained in these operating information, such as the information of various departments, such as the information of various departments. In addition to the system to manage the warehouse management system user information, so to maintain a record of the user operation of the data table, the user's user name, password and operation.

In this theoretical framework, data mining technology is considered as the process of finding the probability distribution of random variables from a large number of source data [9]. For example, Bayesian belief network model, etc.. At present, this method has achieved good results in classification and clustering of data mining. These techniques and methods can be regarded as the development and improvement of the application of probability theory in machine learning. As an ancient subject, statistics has been widely used in data mining.

Association knowledge (Association) is a reflection of an event or an association between an event and an event. Data association in the database is the performance of the real world of things. Database as a structured data form, the attachment to the data model may describe the association between data, such as relational database primary keys and foreign keys. However, the association between the data is complex, not only is the above mentioned in the data model of the association, most of it is hidden.

IV. Design of Intelligent Logistics System Based on Semantic Web and Data Mining Technology

Self built logistics, from the nature of the traditional enterprise logistics network operations, according to the specific needs of customers, the overall balance of the line on the line under the network configuration, based on self built warehousing centers and distribution points to carry out logistics services. In this mode, the supply chain of all aspects of the system and collaboration with high degree, the electricity supplier to grasp the initiative and to carry out targeted improvements to reduce the cost of customer information collection and confidentiality costs, but also to create their own brand image [10]. However, the pattern of capital barrier is high, the need to have a huge amount of orders as support, the enterprise is easy to be excessive expansion of the scale, to a certain extent, affect the company's resource allocation and decision-making, not all electricity suppliers have such a strong financial strength and professional management level.

Definition 4: HTML tags should be in pairs, such as <title></title>, <h1></h1>. But you write HTML file or those by specialized see income development tools automatically generate HTML files, even in grammar mistakes will not affect the HTML file display. Tags in HTML can be individually or not nested, such as <h1><h2></h1></h2>.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} \approx \begin{bmatrix} c'_{11} & c'_{12} & \cdots & c'_{1m} \\ c'_{21} & c'_{22} & \cdots & c'_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c'_{M1} & c'_{M2} & \cdots & c'_{Mm} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix}, \quad C > 0 \qquad (7)$$

As an electronic commerce logistics link, the most important feature is the timeliness, convenience and safety. Electronic commerce logistics link is based on the enterprise computer system's instruction completes the commodity distribution, the supply, the transportation entire process. Intelligent logistics system of the Internet of things is to run in the area of transport vehicle location, delivery of goods type, quantity of management and control. The logistics center is connected to the mobile communication network through the gateway, and the mobile communication network through the M2M and the transport vehicle. Through the GPS system, with the logistics center of the display, the management staff through the GIS maps to easily grasp the current position of the transport vehicle.

Logistics has been the "bottleneck" of the development of e-commerce, with the development of electronic commerce in recent years, the gap between the two. According to relevant statistics, the domestic e-commerce development speed is 200 - 300%, while the logistics growth rate of only 40%, the level of logistics development far can not meet the needs of the development of electronic commerce, especially during the holidays; express Logistics Company appeared frequently "explosion" phenomenon. Coupled with the logistics service level is not high, the arrival of the goods, goods lost, damaged goods, delivery is not in place and other services, has become one of the main complaints of consumers.

RDF is used to provide a simple way of publishing a statement about the Web resources (such as web pages). This section describes the basic idea behind the RDF to provide these capabilities (the specification of these concepts is the RDF concept and abstract syntax [RDF-CONCEPTS]).

This system is mainly divided into the following sub modules: login form module, storage module, a single edit module, a single audit module, the audit of the storage module, employee management module, department management module, the department leader in detail table module and operator set up module, as is shown by Fig .1.

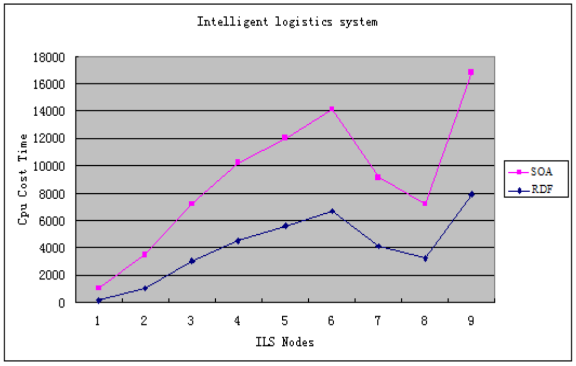


Figure 1. Design of intelligent logistics system based on Semantic Web and data mining technology RDF with SOA.

Data processing, data center to receive come through the Monternet and Internet transmission of data, generate standard database interface, business data storage, analysis, and as a basis for management and decision making; data query. In the end, the holder of the mobile phone will be able to query the various business data, and to deal with various kinds of information management in the enterprise management, realize the functions of receiving and reply, using wireless network and personal affairs management to achieve joint office functions.

V. SUMMARY

E-commerce will promote the improvement of logistics infrastructure, logistics technology and logistics management level. The characteristics of high efficiency and global demand are for electronic commerce requires the improvement of logistics infrastructure, but also to improve the level of logistics technology to improve the efficiency of logistics. In addition, the level of logistics management directly determines the level of logistics efficiency, but also affects the realization of the high efficiency of e-commerce. Only by improving the management level of logistics, establishing scientific and reasonable management system, using scientific management means and methods in logistics management, we can ensure the smooth flow of logistics, realize the rationalization and efficiency of logistics, and promote the development of logistics in electronic commerce.

REFERENCES

[1] SunYu Jie, Julian Clive, "Research of Logistics Product Intelligent System Distribution Based on Internet of Things", JDCTA, Vol. 7, No. 6, pp. 979 ~ 986, 2013

[2] Suphachoke Sonsilphong, Ngamnij Arch-int, "Semantic Interoperability for Data Integration Framework using Semantic Web Services and Rule-based Inference: A case study in healthcare domain", JCIT, Vol. 8, No. 3, pp. 150 ~ 159, 2013.

[3] Krich Intratip, Thepparit Banditwattanawong, Sasiporn Usanavasin, "Stepwise Approach for Applying Coding Method of Grounded Theory to Ontology Design", JCIT, Vol. 8, No. 15, pp. 23 ~ 32, 2013.

[4] Yi Wang, JieHong Luo, "Ontology learning and mapping in semantic web based on formal concept analysis technology", JCIT, Vol. 7, No. 10, pp. 381 ~ 388, 2012.

[5] Putchong Uthayopas, Nunnapus Benjamas, "Impact of I/O and Execution Scheduling Strategies on Large Scale Parallel Data Mining", JNIT, Vol. 5, No. 1, pp. 78 ~ 88, 2014.

[6] Che-Yu Yang, Shih-Jung Wu, "Semantic Web Information Retrieval Based on the Wordnet", JDCTA, Vol. 6, No. 6, pp. 294 ~ 302, 2012.

[7] LI Ning, XU Shoukun , LI Bo, Shi Lin, "An Efficient Ontology-based Semantic Web Services Composition Model for Peer to Peer Work", AISS, Vol. 4, No. 1, pp. 154 ~ 161, 2012.

[8] Yi Wang, JieHong Luo, "Ontology learning and mapping in semantic web based on formal concept analysis technology", JCIT, Vol. 7, No. 10, pp. 381 ~ 388, 2012.

[9] Jinhyung, Myunggwon Hwang, Hanmin Jung, Won-Kyung Sung, "iLaw: Semantic Web Technology based Intelligent Legislation Supporting System", IJIPM, Vol. 3, No. 1, pp. 45 ~ 49, 2012

[10] Che-Yu Yang, Hua-Yi Lin, "Semantic Annotation for the Web of Data: An Ontology and RDF based Automated Approach", JCIT, Vol. 6, No. 4, pp. 318 ~ 327, 2011

Design of Routing Protocol and Node Structure in Wireless Sensor Network based on Improved Ant Colony Optimization Algorithm

Yan Song*
Henan Forestry Vocational College
Henan Luoyang, 471002, China
E-mail: eduhuangweiye@163.com
*The corresponding author

Xiaomei Yao
Henan Forestry Vocational College
Henan Luoyang, 471002, China
E-mail: eduhuangweiye@163.com

***Abstract*—Wireless sensor network is composed of many wireless sensor nodes with the same or different functions. A typical sensor node consists of four parts: sensor unit, information processing unit, wireless communication unit and energy supply unit. In this paper, the existing ant colony algorithm is analyzed, and an improved ant colony optimization algorithm is proposed. The paper presents design of routing protocol and node structure in wireless sensor network based on improved ant colony optimization algorithm. The experimental results show that the proposed method is effective.**

Keywords-Wireless sensor network; Ant colony algorithm; Routing protocol; Optimization; Ant colony system

I. INTRODUCTION

The sensor only as part of a survey project to analyze and study, with the development of material science, and it is especially in 1980s after the development of computer technology and chip integration level, the sensor technology also increased and developed. The sensor is not only used in the parameter measurement range of industrial automatic control, working environment and working medium, and sensor technology and computer technology, the formation of multi function and intelligent micro sensor, makes it easier to popularization and development, control of mobile equipment. Therefore, the extensive use of sensor technology in engineering machinery and it is equipment in order to improve the technology, the performance of these devices.

Wireless sensor networks should meet the following requirements: low energy consumption [1]. The requirements of low energy consumption based on two reasons: one is because the sensor node has the advantages of small volume, so the energy supply is limited; two is due to sensor network work environment is often difficult to update or not because of battery operation cost update. The energy consumption of nodes has a significant impact on the lifetime of wireless sensor networks.

Based on these rules, the ant colony algorithm constructs a heuristic algorithm for the optimal path search using swarm intelligence. Compared with other heuristic search algorithm, to solve the NP complete combinatorial optimal optimization problem, the ACA in the evolution algebra is reduced, the quality of the solution is improved, the convergence speed and solution quality balance in a certain extent. However, the complexity of the analysis shows that the number of ants and the scale of the problem are similar to the number of ants will increase convergence.

At present, many new algorithms have been proposed for the application of ant colony algorithm in wireless sensor network routing [2]. In some literatures, an ant colony algorithm for Steiner tree is proposed, which can be transplanted into WSN routing. However, there is no change in the specific requirements of the WSN, and no consideration of energy consumption is essential to the performance of WSN. It has studied three kinds of ant based WSN algorithms. However, the author only pays attention to the establishment of the initial distribution of pheromone, which has some advantages in the efficiency of the system.

Wireless sensor network is composed of many wireless sensor nodes with the same or different functions. Each sensor node is composed of data acquisition module (sensors, AID converter), data processing and control module (microprocessor, memory), communication module (wireless transceiver) and power supply (battery module, AC/DC energy converter). The node can act as a data collector, a data transfer station or a cluster head, and a cluster head node in the network. With the emergence of wireless sensor networks and the popularity of large-scale, it is possible to obtain the node data efficiently and randomly, but also can avoid the environmental damage caused by the data collection. Wireless sensor networks can be deployed in a large number of sensor nodes in the monitoring area, such as the spread of aircraft, which is practical and convenient, high reliability of the collected information. The paper presents design of routing protocol and node structure in wireless sensor network based on Improved Ant Colony Optimization Algorithm.

II. WIRELESS SENSOR NETWORK NODE ARCHITECTURE AND ROUTING PROTOCOL ANALYSIS

Wireless sensor network (WSN) is a kind of intelligent autonomous measurement and control network system based on the combination of wireless communication technology, sensor technology and network technology. Because of its random layout, the characteristics of self organization, to adapt to the environment, very suitable for wiring, difficult power supply area, inaccessible areas, has been widely used in the field of military defense, industrial and agricultural production, environmental science, traffic management, disaster monitoring, etc..

The hierarchical network communication protocol of wireless sensor network includes physical layer, data link layer, network layer, transport layer and application layer. The physical layer is responsible for data sampling, signal modulation, transmitting and receiving, transmission is responsible for bit stream; data link layer is responsible for the implementation of monitoring data into frames, frames, medium access control, error control, to reduce conflict transmission between the nodes; network layer service.

Sensor node is an important part of wireless sensor network. A typical sensor node consists of four parts: sensor unit, information processing unit, wireless communication unit and energy supply unit. The sensor unit is responsible for the information in the monitoring area and the data of A / D conversion; information processing unit is responsible for the control of the sensor node operation, storage and processing of data and the data itself sent to other nodes; wireless communication unit for wireless communication with other sensor nodes, exchange of control information and data transceiver; energy supply the operation unit provides the energy required for sensor node [3].

When nodes and nodes are deployed in the sensor nodes used for environmental monitoring, it is necessary to choose a small volume, high precision and long life cycle as the monitoring node. As far as possible to reduce the volume of sensor nodes and it is the use of heterogeneous sensor nodes in order to adapt to the complexity of the environment and the sensitivity of the monitoring environment for external equipment. High precision sensor is more conducive to the accurate acquisition of environmental parameters. Another important factor to select the sensor is the start time [4]. The start time is the time between the sensor and the stable read data. The startup time is too long to consume a large amount of energy, which is not conducive to the continuity of the sensor nodes. Therefore, it is necessary to select the sensor with short startup time to save energy, as is shown by equation (1)

$$\begin{cases} a = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^{\,2} - (\sum_{i=1}^{n} x_i)} \\ b = \dfrac{1}{n}\sum_{i=1}^{n} y_i - \dfrac{a}{n}\sum_{i=1}^{n} x_i \end{cases} \tag{1}$$

Data fusion is an important field of integrated intelligent sensor theory, but also the focus of research, data fusion technology, in short, that is to carry out comprehensive treatment of multiple sensors or multi-source information, so as to obtain more accurate and reliable conclusions. The array consists of a plurality of sensors, data fusion technology can give full play to the characteristics of each sensor, using the complementary and redundancy information, improve the measurement accuracy and reliability, prolong the service life of the system.

Wireless sensor networks are starting from the sensor network; the sensor network has experienced the development process. The first generation of sensor networks appeared in 1970s. The traditional sensor with simple information signal acquisition capabilities, using point-to-point transmission, sensing controller connected sensor network; the second generation of sensor networks, with the comprehensive ability, access to a variety of information on the signal, and the interface (such as Rs-232, RS-485) and sensor controller connected to form a sensor network with a comprehensive variety of information; third generation of sensor networks in the late 1990s and the beginning of this century, a variety of sensors to obtain information signal with intelligence, using field bus connecting the sensor controller, constitutes a local area network, become intelligent sensor network.

The serial clock can be a continuous clock to transmit all data in a continuous burst. On the other hand, it can also be a discontinuous clock that sends information to a small number of data to AD7705. DIN pin is the serial data input. The serial data written to the on-chip input shift register is input. The data in the input shift register is transmitted to the setting register, the clock register, or the communication register according to the register selection bit in the communication register. DOUT pin is the serial data output. The output of the serial data read from the on-chip output shift register.

Dynamic characteristics are the characteristics of the output of the sensor when the input changes. In practice, the dynamic characteristics of the sensor are often used to represent the response of some standard input signals. This is because the response of sensors to the standard input signal easily obtained by experimental method, and the relationship between its response to the standard input signal and the input signal of any response, as is shown by equation(2), the former can often know the presumption of the latter. The most commonly used standard input signal has two kinds of step signal and sine signal.

$$E[(X(k)-G(k))^2] = \sum_{i=1}^{N} \left| \frac{u_i}{\sum_{i=1}^{N} u_i} \right|^2 R_i \tag{2}$$

The perception module is responsible for collecting and monitoring data and data conversion processing: processing module is the central node, responsible for data processing, communication network, with power management and positioning and other advanced services; wireless communication wireless communication module is responsible for network nodes, exchange control messages and receiving and sending data: the identity module is responsible for network node identification in number; the power supply module is responsible for the node.

With the miniaturization of the sensor node, the energy of the battery can be limited, and it is difficult to replace the battery because of the unstable environment. So how to limit the energy consumption of nodes becomes the bottleneck of network design, which directly determines the life of the network. On the other hand, the storage capacity of sensor nodes is small, and the computational complexity is low. As a result, researchers in wireless sensor networks pose a challenge for them to design simple and efficient routing protocols for wireless sensor networks.

The communication and networking layer is responsible for point-to-point, point to multipoint wireless communication as well as ad hoc networks, and provides service support to management and basic service layer. The network communication protocol of physical layer, data link layer and network layer of wireless sensor network is studied in this paper. The main problem of physical layer is the choice of wireless frequency band and modulation technology. The research focus of the data link layer is the media access control, which is the method of allocating channel resources among competing users.

Compared with the limited processing power, storage capacity and communication ability of sensor nodes, and it is sink nodes have strong processing power, storage capacity and communication ability. The sink node is directly connected to the sensor node network and communication network, which brings forward higher requirements for the functionality of the sink node. The architecture of a typical wireless sensor network protocol stack to support dynamic and hierarchical, adaptive, programmable, self management, self recovery, multi task and proxy based it.

Modeler modeling process is divided into 3 levels: process (process) level, node (Node) level and network (Network) level. The behavior of a single object is simulated at the process level, and it is connected to the device at the node level. Several different network scenarios form a "project" to compare different design scenarios. This is also an important mechanism for Modeler modeling; this mechanism is conducive to project management and division of labor.

A high speed digital signal processor (DSP) TMS320F240 is used as the central processing unit, and the core control unit is composed of a few peripheral circuits. The main advantages of this scheme: the speed is fast, the execution speed is 20MIPS, almost all of the instructions can be completed in a single cycle of 50ns, and such a high performance is very suitable for real-time data acquisition [5].

$$H(x,y)=\begin{bmatrix}\frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x\partial y}\\ \frac{\partial^2 f}{\partial y\partial x} & \frac{\partial^2 f}{\partial y^2}\end{bmatrix} \tag{3}$$

Typically, most sensor nodes are stationary. In addition, the energy, processing power, storage capacity and communication ability of sensor nodes are very limited. The primary design goal of traditional wireless network is to provide high quality service and high bandwidth utilization, then considering energy saving; and the primary design goals of sensor networks is the efficient use of energy, which is one of the most important difference between sensor networks and traditional networks.

The static characteristic of the sensor is the relationship between the input signal and the output of the sensor. When the input and output are independent of time, so the relationship between them, namely sensor static characteristics available a time-dependent algebraic equation, or input as abscissa, the output characteristic curve and the corresponding longitudinal coordinate and draw to describe. The main parameters that characterize the static characteristics of the sensor are linearity, sensitivity, hysteresis, repeatability, drift and so on.

Passive routing considers that there is no need to maintain routing information for all nodes in a dynamically changing network, and it will only be "on-demand" when there is no destination node routing. According to the request of network transmission, the passive node searches from the source node to the destination node.

III. Improved Ant Colony Optimization Algorithm

The ant cycle model, which is used by the traditional basic ACA, adopts the ant model which is more close to the real ant behavior. The establishment of pheromone diffusion model, the closer distance between the ants can collaborate better, the simulation results show the effectiveness of the proposed algorithm in this paper is however required to achieve convergence in evolutionary algebra is the basic of ACA has been greatly improved. Reduce about 4 times [6]. However, the reduction of the shortest path length is not obvious, and the setting of the parameters is still based on the experiment.

The individuals in the group are distributed (Distributed) so that they can adapt to the working conditions in the current network environment. (2) There is no central control and data, such a system is more robust (Robust), not because of the failure of one or a few individuals affect the whole problem solving. (3) Such a system can be more scalable (Scalability) without direct communication between individuals, rather than through direct communication.

Based on the order of the ant system (Rank-based Version Ant System, referred to as RAS) is a AS Bullnheimer and other extensions of the algorithm proposed by Bernd. RAS after it is each iteration, the ant path will arranged according to the order from small to large, namely L1 (T) = L2 (t)......... Lm (T), and according to the length of the path with different weights, shorter path length the greater the weight. The weight of the global optimal solution is w, and the weight of the R optimal solution is max {0, w-r}, and update the information of each path according to (4) [7].

$$H_e=-\sum_{l=0}^{L-1}P(l)\log_2 p(l) \tag{4}$$

The research of ACA is more and more deep, all kinds of model ACA based on the research domain also emerge in an endless stream, scattered from the domain to the continuous domain, at the same time ACA is combined with other algorithm in order to overcome the defects of the domestic research starts late, to affect the convergence of the parameters such as B, has been unable to determine a set of related the theory to be set, only through repeated tests to determine approximately a range of parameters, and the research about the theory simulation, applied to practice is still less. The study of these areas abroad has been more mature.

The pheromone values prescribed in the range; third, the initial pheromone value is 7, the combination, the purpose is to make the algorithm to explore more unknown; finally, when the algorithm premature convergence phenomenon, or cycle after a certain number of algorithms still did not find the shorter path, then all the information on the path each will be reset back to the initial state when looking into the current optimal path, the algorithm will update according to the global information system in the current rules of the ant optimal path pheromone update.

The idea of ACO was proposed in 1991 by Italy scholar M.Dorigo et al. From 1991 to 1996, M Dorigo et al. Ant colony search process and traveling salesman problem (TSP) food similarity, do some research by artificial ants search for food in the process, has put forward three kinds of models: ant-quantity, ant-density and ant-cycle. The main difference between the three models lies in the different mathematical formulas for the change of the pheromone concentration. M Dorigo published a comprehensive discussion on the ant system (AS), summed up the three models [8]. In this paper, the M model is introduced into ant-code Dorigo, and a series of experiments are done for the TSP problem.

$$C = \sum_{i=0}^{L-1} p_i \log_2 \frac{p_i}{q_i} \tag{5}$$

Set the cycle counter, set the initial amount of information for each path, and place the ants on a single city. The set of index set, from 1 to the ants in the starting position, set the corresponding set repeat the following steps, until the set date (this step will be repeated): setting; for from 1 to 4, according to the probability formula to determine the choice of target moving under a city step.

Ant colony optimization is based on ant colony optimization. Based on the disadvantages of AS optimization ranking elitist AS, through sorting can well restrain premature, especially when the initial state of the solution had little difference, but the effect is significant, which increase the amount of pheromone path optimal ant induced on the third.

From the perspective of the overall analysis process of ACRP, we need to empty all the ants list, initialization pheromone path value; then began to traverse the source node around the adjacent node, according to the path calculation of optimal selection probability, in order to continue to visit the next node, update the pheromone path and the list, prevent the entry of death cycle, until you reach the Sink node; Sink nodes at the optimal path, the global update pheromone on the optimal path of information; if still not many times to find a better solution to the pheromone evaporation rate update.

IV. Design of Routing Protocol and Node Structure in Wireless Sensor Network Based on Improved Ant Colony Optimization Algorithm

Wireless sensor network (WSN) is a new information acquisition technology with the development of wireless communication technology, sensor technology, microelectronics technology and distributed information processing technology after Internet. WSN combines the embedded technology, sensor technology, communication technology and distributed information processing technology, to various environmental cooperation real-time sensing, monitoring and collecting network of regional distribution of information, and the data were processed to obtain appropriate simplification and accurate information, and send to the end user.

In sensor networks, the use of the process, some sensor nodes due to energy depletion or environmental factors caused by the failure, there are also some nodes in order to compensate for the failure of nodes and increase the monitoring accuracy and added to the network, so that the number of nodes in a sensor network can dynamically increase or decrease, so that the topology of the network with dynamic changes. The self-organization of sensor networks should be able to adapt to the dynamic changes of the network topology.

Ant colony system (AS) is with elitist strategy. In the elitist AS, in order to find the optimal solution so far in the next cycle of ants are more attractive, give the optimal solution additional amount of pheromone update it is said by the ant pheromone induced by pheromone path increasing after each iteration; the number is elite ants; path length is optimal to find solutions [9]. This algorithm has fast convergence speed and short computation time, but if it is too large, the search will be limited to the extreme value, as is shown by equation (6).

$$\overline{R}_i(k) = \frac{1}{k}\sum_{j=1}^{k} R_i(j) = \frac{1}{k}\left[\sum_{j=1}^{k-1} R_i(j) + R_i(k)\right] \tag{6}$$

Hierarchical routing is a routing protocol. Hierarchical routing protocol uses cluster to classify sensor nodes, which is the concept of adjacent node clusters. The cluster head is used to complete the communication in the cluster, and the cluster head node collects the information of the nodes in the cluster, so as to reduce the traffic volume, and finally, the collected data is transmitted to the terminal node through the cluster head node [10]. Therefore, the wireless sensor network is scalable, while avoiding the energy consumption of sensor nodes to prolong the network lifetime.

In the study, the network is often abstracted as weighted directed graph G (V, A). Where V is the node set, the V element of graph nodes; a arc set, a elements (I, J) and A V (I) for node to node V (J) of arc; arc is a weighted metric, metric operations include additive, multiplicative and minimum. Business requirements for QoS metrics are QoS constraint C= (C1, C2,,, CK). QoS routing is found in figure G, the (0) to V (V) (V) (V) (P=v) (n) (n) is found to satisfy the service QoS constraint C, from the source node to the target node (2). It should meet the quality requirements of different services and the effective utilization of resources in the whole network.

If there is no record of the ant in the node memory, the required information is saved, and the node visited by the ant is updated. If the serial number of the ant is found in the node memory, the ant will be deleted. When the node receives an ant returning is to the source node, and it is the node searches for the last node that is in the forward direction, which is regarded as the next node. If the ant does not return to the node for a period of time due to external factors, the corresponding information of the ant will be removed.

So the features and advantages of using ant colony optimization proposed a ZigBee routing algorithm based on Ant Colony Optimization Based on ZigBee routing strategy and ant colony optimization features, constructing artificial ants algorithm has the advantages of energy saving and better global optimization ability, and improve the performance of the network. Put forward its own ideas and practices to achieve the purpose of network optimization.

V. SUMMARY

Wireless sensor networks (WSN) are task oriented networks, and it is meaningless to talk about sensor nodes from the sensor network. The nodes in the sensor network are identified by the number of nodes, and whether the node number needs the whole network is decided by the design of the network communication protocol. Due to the random deployment of sensor nodes, the relationship between the sensor network and the node number is completely dynamic. The paper presents design of routing protocol and node structure in wireless sensor network based on Improved Ant Colony Optimization Algorithm. When the ant arrives at the node on the path, it collects the information of the pheromone above, and when it reaches the destination node, it updates the pheromone table with the information. Then, the destination node will send back the ant, whose task is to update the pheromone table of the source node.

REFERENCES

[1] Christian Blum Ant Colony Optimization Introduction and Recent Trends[J].Physics of Life,2005(10):10-20

[2] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson, “Wireless sensor networks for habitat monitoring,” in Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications. ACM Press, 2002, pp. 88–97.

[3] IRIDIA.Ant Colony 0ptimization − − Artificial Ants as AComputational Intelligenee Technique[J].IRIDIA - Techinal Report Series,2006(6):115-126.

[4] Aline,Baggio.Wireless Sensor Networks in precision agriculture.REAL WSN 2005, Workshop on RealWord Wireless Sensor Networks,Storkholm,Sweden,June 2005:20-21.

[5] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “A survey on sensor networks,” IEEE Communications Magazine, vol. 40, no. 8, pp. 102–114, August 2002.

[6] Z. Butler, D. Rus. Enent-based motion control for mobile-sensor networks. Pervasive Computing, 2003, 2(4): 34-42.

[7] W. Ye, J. Heidemann, and D. Estrin, “An energy efficient MAC protocol for wireless sensor networks,” in In proceedings of the 21st Conference of the IEEE Computer and Communications Societies (INFOCOM), vol. 3, June 2002, pp. 1567–1576.

[8] B Bullnheimer, R F Hartl, and C. C Strauss. A new rank-based version of the Ant System: A computational study [J]. Central European Journal for Operations Research and Economics, 1999, 7(1):25-38.

[9] J. Kulik, W. Heinzelman, and H. Balakrishnan, “Negotiation-based protocols for disseminating information in wireless sensor networks,”Wirel. Netw., vol. 8, no. 2/3, pp. 169–185, 2002.

[10] Chong CY, KumarS. Sensor networks: Evolution, opportunities, and challenge [J]. Proceedings of the IEEE, 2003,91-102

Forecasting CSI 300 index using a Hybrid Functional Link Artificial Neural Network and Particle Swarm Optimization with Improved Wavelet Mutation

Tian Lu
Dept. of Math. and Physics
North China Electric power university
Beijing, China
E-mail: lutian515@163.com

Zhongyan Li
Dept. of Math. and Physics
North China Electric power university
Beijing, China
E-mail: lzhongy@ncepu.edu.cn

***Abstract*—Financial market dynamics forecasting has long been a focus of economic research. A hybridizing functional link artificial neural network (FLANN) and improved particle warm optimization (PSO) based on wavelet mutation (WM), named as IWM-PSO-FLANN, for forecasting the CSI 300 index is proposed in this paper. In the training model, it expands a wider mutation range while apply wavelet theory to the PSO, in order to exploring the solution space more effectively for better parameter solution. In the stimulating experiment, we use five benchmark functions to test the proposed method, and the results shows that IWM-PSO has greater convergence accuracy than WM-PSO and PSO. The empirical research is performed in testing the predictive effects of CSI 300 index in the proposed model compared with the back propagation functional link neural network (BP-FLANN), PSO-FLANN and WM-PSO-FLANN. The experiment utilizes two expansion functions, Chebyshev functions and trigonometric functions, to map the input data to higher dimension. The results show that the prediction performance of the proposed model displays a better performance in financial time series forecasting than other three models. Moreover, the accuracy of the input with trigonometric functions is higher, and it suggests that trigonometric function is more suitable for this kind of data type.**

Keywords-Stock index; Forecasting; Functional link artificial neural network (FLANN); Improved wavelet mutation (IWM); Particle warm optimization (PSO)

I. INTRODUCTION

The CSI 300 index is composed of the 300 largest and most liquid A-shares listed on the two stock exchanges, launched on April 8th, 2005, and it aims to reflect the price fluctuation and performance. The CSI index has been used as the basis for many financial products around the world and is also used by investors to develop and benchmark their portfolios [1].

Predicting the stock market index is of prime importance in private and institution investors when making investment decision because successful prediction of stock prices may be guaranteed benefits. Their mainly quest is to forecast, to determine the appropriate time to buy, hold or sell. The investors assume that the future trend of the stock market is based at least in part on present and past events and data. However, the price variation of stock market is a dynamic system and the disordered behavior of the stock price movement duplicates complication of the price prediction and the highly non-linear, dynamic complicated domain knowledge inherent in the stock market makes it very difficult for investors to make the right investment decisions. Many researchers in the past have applied various computing techniques to predict the movement of the stock markets.

Artificial neural network have gained its popularity due to their inherent capability to approximate any none-linear function, less sensitive to error, tolerate noise, and chaotic components. To improve predicting precision, various network architectures and learning algorithms have been developed in the papers [2-4].

Further it is well known that the artificial neural network (ANN) suffers from local minima trapping, saturation, weight interference, initial weight independence and over fitting, make ANN training difficult. Additionally, it is also very difficult to fix the parameters like number of neurons in a layer, and number of hidden layers in a network, thereby deciding a proper architecture is not that easy. Thus to overcome from this functional link artificial neural network (FLANN) based simple network may be used for the prediction the stock index data with less computational overhead than the ANN [5]. FLANN is proposed by Y.H.PAO and Y. Takefji [6] contains a single layer neural network in which nonlinearity is introduced as a functional block, thus giving rise to higher dimension input space. In FLANN, higher-order correlations among input components can be used to construct a higher –order network to perform non- linear mapping using only a single layer of units.

The researchers proposed different FLANN models to predict the stock market indices, such as DJIA, S&P 500 stock indices[7,8] Indian stocks[9,10], and there is no research about the CSI 300 index.

Particle swarm optimization (PSO) is a population-based stochastic optimization algorithm which inspired by the social behaviors of animals like fish schooling and bird flocking proposed by Kennedy J and Eberhart R [11]. The PSO has comparable or even superior search performance for many hard optimization problems with faster and more stable convergence rate. The PSO has been widely used in parameter learning of neural network [12-14]. We can harness the power of PSO to train our model to reduce the possibility to be trapped in local optimal and speed up the convergence.

The rest of the paper is organize as follows: in section 1, we will give a brief introduction of the study. Section 2 deals with model development using FLANN structure using the improved wavelet mutation based PSO algorithm. The details experiment setup including process are discuss in Section 3. The simulating study and the results obtained from experiment are provided in section 4. In the final section the conclusion has been made.

II. Model development

A. The FLANN Model

Depending the complexities of the problems, number of layers and number of neurons in the hidden layer need to be changed. As the number of layers and the number of neurons in the hidden layer associated with multi-layer neural network increases, training the model becomes more complex. In FLANN, each element of the input data undergoes functional expansion through a set of basis functions to enhance the input pattern with nonlinear functional expansion. This enables the FLANN to solve complex problems by generating non-linear decision boundaries.

The architecture consist of two parts, namely, transformation part and learning part. The transformation deals with the input feature vector to hidden layer by approximate transformable method. A simple FLANN model with a pattern of two features is shown in Fig .1.

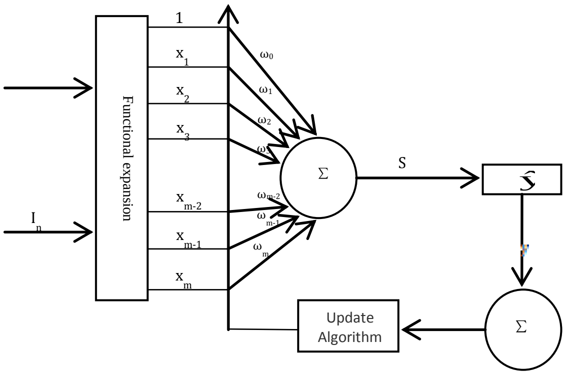


Figure 1. Structure of FLANN model with a pattern of two features

Let us consider a two dimension input sample $I = [I_1, I_2]$. This sample is mapped to a higher dimensional space by functional expanding using trigonometric functions (1):

$$X = [(I_1, \cos\pi I_1,, \sin\pi I_1 \cdots \cos N\pi I_1,, \sin N\pi I_1), \\ (I_2, \cos\pi I_2,, \sin\pi I_2 \cdots \cos N\pi I_2,, \sin N\pi I_2)]^T \tag{1}$$

Each input sample is expanded to N sine Terms, N cosine terms plus the sample itself. If Chebyshev polynomial basis function (2) is employed

$$X = [(Ch_0(I_1), Ch_1(I_1) \cdots Ch_{n+1}(I_1)), \\ (Ch_0(I_2), Ch_1(I_2) \cdots Ch_{n+1}(I_2))]^T \tag{2}$$

where $Ch_0(x) = 1$, $Ch_1(x) = x$, $Ch_{n+1}(x) = 2xCh_n(x) - Ch_{n-1}(x)$, and n is the order of the polynomial chosen.

The output of hidden neuron is given:

$$\hat{y} = \tanh(\sum_{i=1}^{m} x_i w_i - \theta) \tag{3}$$

Where θ is the threshold of neuron in the output layer.

B. Improved Wavelet Mutation Based Particle Swarm Optimization

In [15], it proposed an improved version of the PSO, where the constriction inertia weight factors are introduced, the velocity v_{ij}^t and the position x_{ij}^t of the element the *j*th of the particle *i*th at the *t*th iteration can be calculated using (4)(5):

$$v_{ij}^{t+1} = k\left((wv_{ij}^t + \phi_1 r_1)\left(pbest_{ij}^t - x_{ij}^t\right) + \phi_2 r_2\left(gbest_j^t - x_{ij}^t\right)\right) \tag{4}$$

$$x_{ij}^{t+1} = x_{ij}^t + v_{ij}^{t+1} \tag{5}$$

where $pbest_I^t$ is the best position in the history of particle *i*th at the *t*th iteration; $gbest^t$ is the global best position in the history at the *t*th iteration; r_1 and r_2 are uniform random number in the range of [0,1]; φ_1 and φ_2 are acceleration constants is a constriction factor derived from the stability analysis of to ensure the system to be converged but not prematurely. Mathematically, *k* is a function of φ_1 and φ_2 as reflected in (6):

$$k = \frac{2}{\left|2 - \varphi - \sqrt{\varphi^2 - 4\varphi}\right|} \tag{6}$$

Where $\varphi = \varphi_1 + \varphi_2$, and $\varphi > 0$

w is inertia weight. Generally can be dynamically set with (7):

$$w = V\max - \left(V\max - V\min\right)\frac{t}{T\max} \tag{7}$$

Where *t* is the current iteration number, *Tmax* is the total number of iteration, and *Vmax* and *Vmin* are the upper and lower limits of the inertia weight.

However, observation reveals that the PSO sharply converges in the early stages of the searching process, but saturated or even terminates in the later stages. In [11], it proposed the operation of the hybrid PSO with a wavelet mutation. Every particle element of the swarm will have a chance to mutate that is governed by a probability of mutation p_m, which is defined by the user. For each particle element, a random number between 0 and 1 will be generated such that if it is less than or equal to p_m, a mutation will take place on that element.

The element of particle is randomly selected for the mutation (the value of x_{ij}^t is inside the particle element's boundaries [$para_{\min}^j$, $para_{\max}^j$]), the resulting particle is given by (8):

$$\mathrm{mut}(x_{ij}^t)=\begin{cases} x_{ij}^t+\sigma*(para_{\max}^j-x_{ij}^t),\sigma>0 \\ x_{ij}^t+\sigma*(x_{ij}^t-para_{\min}^j),\sigma\le 0 \end{cases} \tag{8}$$

The Morlet wavelet integrates to zero. Over 99% of the total energy of the function is contained in the interval of [-2.5, 2.5]. By using Morlet wavelet in as the mother wavelet.

$$\sigma=\frac{1}{a}e^{-\left(\frac{b}{a}\right)^2/2}*\cos\left(5*\frac{b}{a}\right) \tag{9}$$

Hence, the overall positive mutation and the overall negative mutation throughout the evolution are nearly the same. This property gives better solution stability. $\sigma = 1$， $mut(x_{ij}^t)= para_{\max}^j$;$\sigma= -1$， $mut(x_{ij}^t)= para_{\min}^j$. A larger value of $|\sigma|$ gives a larger searching space for fine-tuning. b can be randomly generated form [-0.25a, 0.25a]. A monotonic increasing function governing a and $t/Tmax$ is proposed as follows:

$$a=e^{-\lg(g)*\left(1-\frac{t}{T\max}\right)^{\xi_{\mathrm{wm}}}+\lg(g)} \tag{10}$$

where ξ_{wm} is the shape parameter of the monotonic increasing function, and g is the upper limit of the parameter a.

However, after mutation, the value of the particle element is between [$para_{\min}^j$, $para_{\max}^j$]. If the particle did not go by the best solution, it still has the possibility of being trapped in the local optima. So we propose the particle swarm optimization based on an improved wavelet mutation by expand the mutation range during each generation, so that the particle would be more likely to fly near the global best optima. The equation of the mutation can be changed into (11):

$$\mathrm{mut}(x_{ij}^t)=\begin{cases} x_{ij}^t+\sigma*(\eta_{\max}*para_{\max}^j-x_{ij}^t),\sigma>0 \\ x_{ij}^t+\sigma*(x_{ij}^t-\eta_{\min}*para_{\min}^j),\sigma\le 0 \end{cases} \tag{11}$$

Where $\eta_{\max}$ and $\eta_{\min}$ is defined by users to expand the mutation range.

III. Analysis of Datasets

A. Stimulation Study

TABLE I. BENCHMARK FUNCTIONS

Function expression	ξ_{wm}
$f_1=\sum_{i=1}^{\mathrm{n}} x_i^2$	2
$f_2=\sum_{i=1}^{D}\left(\sum_{j=1}^{i} x_i^2\right)$+random[0,1)	4
$f_3=\frac{1}{4000}\sum_{i=0}^{\mathrm{n-1}} x_i^2+\sum_{i=0}^{\mathrm{n-1}}\cos(\frac{x_i}{\sqrt{i+1}})+1$	4
$f_4=\sum_{i=1}^{\mathrm{n}}(x_i^2\text{-10cos }(2\pi \mathrm{x}_i)+10)$	6
$f_5=0.5+\frac{\sin(\sqrt{x_1^2+x_2^2})^2-0.5}{\left(0.1+0.001\left(x_1^2+x_2^2\right)\right)^2}$	6

A suite of five benchmark test functions is used to test the performance of the proposed model. Many different kinds of optimization problems are covered by these benchmark test functions. These five benchmark functions can be divided into three categories. The first one is the category of the unimodal functions, which is a symmetric model with one single minimum, f_1 and f_2. The second one is the category of multimodal functions with some local minima, f_3 and f_4. The third category is low dimension function, f_5. The benchmark function expressions and their parameter ξ_{wm} setting for different functions are shown in Table I.

B. Empirical Dataset

Empirical study is carried out using the date set CSI 300 from 2008/1/2 to 2017/3/15 up to 2236 trading days. Following 7:3 ratios, we use 1566 days for training and remaining 670 days for validating the model. Choose five kinds of stock prices as the input values in the input layer: daily highest price, daily lowest price, daily closing price, change rate and turnover, and the output layer is the closing price of the next trading day. The entire data is normalized to values between 0 and 1. The normalization formula in (12) to express the data in terms of the minimum and maximum value of the dataset.

$$\mathrm{input}=\frac{\mathrm{x}-Min.}{Max.-Min} \tag{12}$$

Where input and x represents normalized and actual value respectively.

We use both Chebyshev functions and trigonometric functions to map the data into high dimension. In trigonometric functions, we set the N is 2, and the swarm dimension would be 26, the size of the swarm we set is 60. In trigonometric functions, we set the n is 3, and the swarm dimension would be 16, the size of the swarm we set is 40. And the p_{m} is 0.2 and ξ_{wm} we set is 0.2.

To compare the forecasting performance of four considered forecasting models, IWM-PSO-FLANN, WM-PSO-FLANN, PSO-FLANN and BP-FLANN, we use the following criteria: the mean absolute error (MAE)(13), the root mean absolute error (RMSE)(14) and the mean absolute percentage error (MAPE)(15), the corresponding definition are given as follows:

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{t=1}^{N}\left|y_t - \hat{\mathrm{y}}_{\mathrm{t}}\right| \tag{13}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{t=1}^{N}\left(\left|y_t - \hat{\mathrm{y}}_{\mathrm{t}}\right|\right)^2} \tag{14}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{t=1}^{N}\left(\left|y_t - \hat{\mathrm{y}}_{\mathrm{t}}\right|\right)^2} \tag{15}$$

Where y and $\hat{y}$ represents the actual and forecast values; N is the total number of the data. Noting that MAE, RMSE, and MAPE are measures of the deviation between the prediction values and the actual values, the prediction performance is better when the values of these evaluation criteria are smaller. However, if the results are not consistent among these criteria, we choose the MAPE as the benchmark since MAPE is relatively more stable than other criteria.

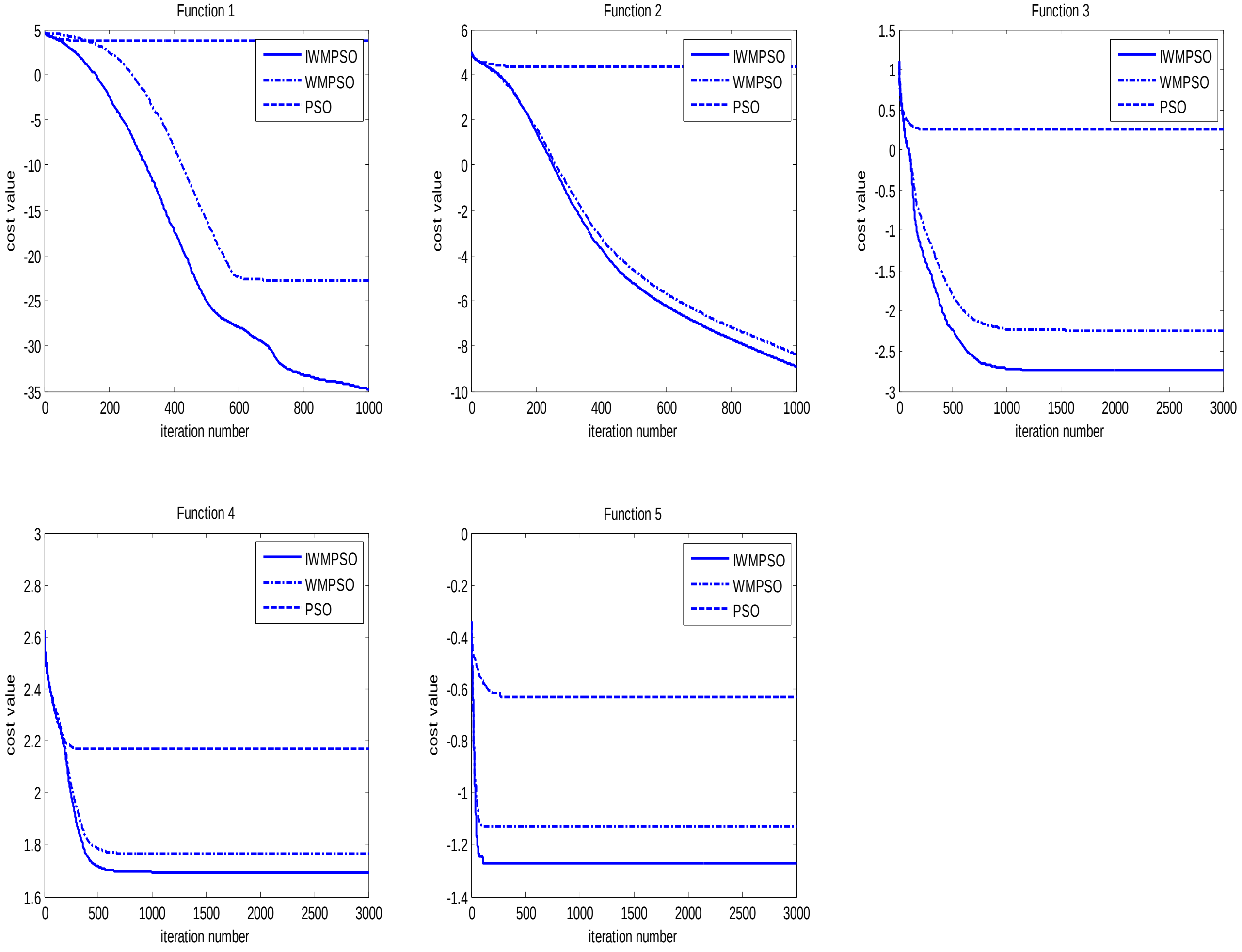


Figure 2. Convergence curves of benchmark function for different PSO method

IV. Results and Dicussion

A. Benchmark Test Functions

The result for the five benchmark test functions are in Table II and Fig .2. From Fig .2, The IWM-PSO will get the same accuracy as other two models in fewer iteration times. And from Table II, results of stimulating experiment in terms of the mean cost values and the best cost value of the IWM-PSO are much better than those of the other methods. Also, the standard deviation is much better, which means that the searched solutions are more stable. From what mentioned below, we can conclude that the IWM-PSO will have better and more stable solution compared to other models.

TABLE II. Comparison between different PSO method for benchmark test functions.

		IWM-PSO	WM-PSO	PSO
f_1	Mean	-34.7321	-22.7162	3.6758
	Best	-47.1855	-38.8609	2.6199
	Std Dev	-68.0077	-44.0165	1.2541
f_2	Mean	-8.8846	-8.3512	4.3211
	Best	-28.8609	-10.5959	3.4875
	Std Dev	-18.9078	-12.5434	0.16147
f_3	Mean	-2.7140	1.7642	2.1664
	Best	1.3407	1.2022	1.9377
	Std Dev	2.3953	2.8338	2.9743
f_4	Mean	1.6944	1.7642	2.1664
	Best	1.3407	1.2022	1.9377
	Std Dev	2.4953	2.8338	2.9743
f_5	Mean	-1.2728	-1.1299	-0.6316
	Best	-3.1565	-2.0125	-1.4325
	Std Dev	-2.0186	-1.7167	-1.3677

B. CSI 300 Index Prediction

Fig .3. Shows the predictive values on the test set for CSI 300 ((a) uses trigonometric functions and (b) uses Chebyshev functions). The empirical research shows that the proposed the IWM-PSO-FLANN model has the best performance. When the stock market is relatively stable, the forecasting result is near to the acturacl value. At the same time, we can see that the large fluctuation period forcasting is relatively not that accurate from these four models. But the IWM-PSO-FLANN has the best performance than other three models in this period. In Table III, the results show that the forecasting results of the proposed model are almost smaller than those by other models and these can conclude that the proposed IWM-PSO-FLANN model is better than the three other models. Morever, the model expaned with trigonometric functions has smaller values of MAE, RMSE and MAPE, which mean that trigonometric functions is more suitable for this kind of data type.

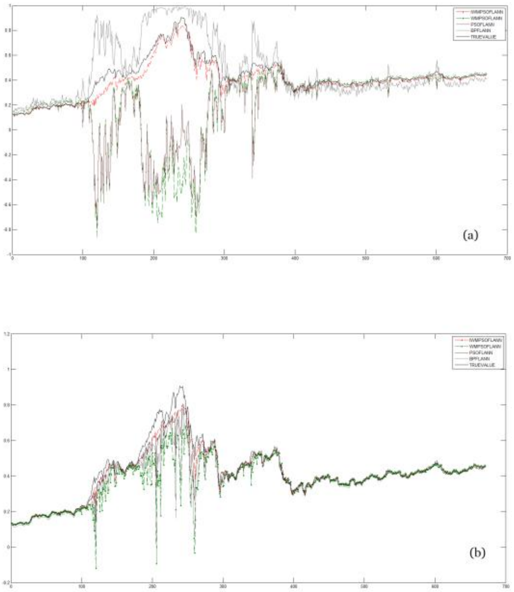


Figure 3. Forecasting result of CSI 300 stock insex for four forecasting models

TABLE III. Comparision of index's prediction for different forecasting model

	trigonometric functions			
	IWM-PSO-FLANN	WM-PSO-FLANN	PSO-FLANN	BP-FLANN
MAE	0.018266	0.050491	0.025934	0.039624
RMSE	0.032258	0.111879	0.050256	0.088087
MAPE	3.766379	9.09154	5.126772	7.784697
	Chebyshev functions			
	IWM-PSO-FLANN	WM-PSO-FLANN	PSO-FLANN	BP-FLANN
MAE	0.028131	0.229882	0.213812	0.098014
RMSE	0.043821	0.456479	0.414253	0.143458
MAPE	6.255943	41.097287	38.053276	39.453523

Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression, “One of us (R.B.G.) thanks . . .” Instead, try

“R.B.G. thanks”. Put applicable sponsor acknowledgments here; DO NOT place them on the first page of your paper or as a footnote.

References

[1] Biktimirov E N, Li B. Asymmetric stock price and liquidity responses to changes in the FTSE Small Cap index[J]. Review of Quantitative Finance and Accounting, 2014, 42(1):95-122.

[2] Lahmiri S. Wavelet low- and high-frequency components as features for predicting stock prices with backpropagation neural networks[J]. Journal of King Saud University - Computer and Information Sciences, 2014, 26(2):218-227.

[3] Rather A M, Agarwal A, Sastry V N. Recurrent neural network and a hybrid model for prediction of stock returns[J]. Expert Systems with Applications, 2015, 42(6):3234-3241.

[4] Oliveira F A D, Nobre C N, Zárate L E. Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index – Case study of PETR4, Petrobras, Brazil[J]. Expert Systems with Applications, 2013, 40(18):7596-7606.

[5] Das S, Patra A, Mishra S, et al. A self-adaptive fuzzy-based optimised functional link artificial neural network model for financial time series prediction[J]. International Journal of Business Forecasting & Marketing Intelligence, 2015, 2(1):55.

[6] Pao Y H. Adaptive pattern recognition and neural networks [M]. Addison-Wesley, 1989.

[7] Majhi R, Panda G, Sahoo G. Development and performance evaluation of FLANN based model for forecasting of stock markets[J]. Expert Systems with Applications An International Journal, 2009, 36(3):6800-6808. A

[8] Rout M, Majhi B, Mohapatra U M, et al. Stock indices prediction using radial basis function neural network[C]// International Conference on Swarm, Evolutionary, and Memetic Computing. 2012:285-293.

[9] Bebarta D K, Biswal B, Dash P K. Comparative study of stock market forecasting using different functional link artificial neural networks[J]. International Journal of Data Analysis Techniques & Strategies, 2012, 4(4):398-427.

[10] Rout M, Majhi B, Mohapatra U M, et al. Stock indices prediction using radial basis function neural network[C]// International Conference on Swarm, Evolutionary, and Memetic Computing. 2012:285-293.

[11] Kennedy J, Eberhart R. Particle swarm optimization [C]//Proceedings of the IEEE International Conference on Neural Networks, Perth, 1995: 1942 – 1948.

[12] Pulido M, Melin P, Castillo O. Particle swarm optimization of ensemble neural networks with fuzzy aggregation for time series prediction of the Mexican Stock Exchange[J]. Information Sciences, 2014, 280:188-204.

[13] Wang S, Zhang Y, Dong Z, et al. Feed-forward neural network optimized by hybridization of PSO and ABC for abnormal brain detection[J]. International Journal of Imaging Systems & Technology, 2015, 25(2):153-164.

[14] Bouacha K, Terrab A. Hard turning behavior improvement using NSGA-II and PSO-NN hybrid model [J]. International Journal of Advanced Manufacturing Technology, 2016:1-20.

[15] Zhao B, Guo C X, Cao Y J. A multiagent-based particle swarm optimization approach for optimal reactive power dispatch"[J]. IEEE Transactions on Power Systems, 2005, 20(2):1070-1078.

[16] Ling S H, Yeung C W, Chan K Y, et al. A new hybrid Particle Swarm Optimization with wavelet theory based mutation operation[C]// Evolutionary Computation, 2007. CEC 2007. IEEE Congress on. 2007:1977-1984.

Equivalence on Quadratic Lyapunov Function Based Algorithms in Stochastic Networks

Li Hu
National Key Laboratory of Science and Technology on
Test Physics and Numerical Mathematics
Beijing, 100076, P. R. China
E-mail: sdbzlh@163.com

Liu Jiaqi
National Key Laboratory of Science and Technology on
Test Physics and Numerical Mathematics
Beijing, 100076, P. R. China
E-mail: sdbzlh@163.com

Gao Lu
National Key Laboratory of Science and Technology on
Test Physics and Numerical Mathematics
Beijing, 100076, P. R. China
E-mail: sdbzlh@163.com

Wang Shangyue
National Key Laboratory of Science and Technology on
Test Physics and Numerical Mathematics
Beijing, 100076, P. R. China
E-mail: sdbzlh@163.com

***Abstract*—Quadratic Lyapunov function based Algorithms (QLAs) for stochastic network optimization problems, which are cross-layer scheduling algorithms designed by Lyapunov optimization technique, have been widely used and studied. In this paper, we investigate the performance of using Lyapunov drift and perturbation in QLAs. By analyzing attraction points and utility performance of four variants of OQLA (Original QLA), we examine the rationality of OQLA for using the first-order part of an upper bound of Lyapunov drift of a function L_1. It is proved that either using the real Lyapunov function (L_2) of networks under QLA or using the entire expression of Lyapunov drift does not improve backlog-utility performance. The linear relationship between the attraction point of backlog and perturbation in the queue is found. Simulations verify the results above.**

Keywords-Component; Lyapunov optimization; QLA; Lyapunov function; Backlog-utility performanc; Stochastic network optimization

I. INTRODUCTION

Lyapunov optimization technique is an effective method to design online cross-layer scheduling algorithms for stochastic network. The Lyapunov optimization technique is able to stabilize the network while achieving close-to-optimal utility performance [1][2]. Among the multiple advantages of using the Lyapunov optimization technique in stochastic network optimization, the most significant one is that probability distributions in the network are not necessarily known but able to be obtained by the Lyapunov optimization technique, adapt to networks with any distributions. The Lyapunov optimization technique has been used in various scenarios, including wireless communication networks [3][4], energy harvesting networks [5], processing networks [6], and even financial systems [7].

We mainly focus on the Quadratic Lyapunov function based Algorithm (QLA). For clarity, the original QLA proposed in e.g. [3] is referred as OQLA henceforth. OQLA is designed to greedily minimize an expression consisting of two parts, one of which is the first-order part of the upper bound of Lyapunov drift of a specific Lyapunov function (this function is denoted as L_1), e.g. [2]. However, using the entire expression of Lyapunov drift when minimizing seems to perform better than using first-order part of upper bound of Lyapunov drift. Moreover, L_1 used in design OQLA is apparently not the actual Lyapunov function of networks under OQLA. Because if L_1 is the actual Lyapunov function of networks under OQLA, queue backlog should be attracted by zero. But both analysis (e.g. in [8]) and simulation (e.g. in [2]) show that queue length is attracted by a non-zero value in networks under OQLA. It seems quite arbitrary to design OQLAs by using L_1 and the first-order part. To our best knowledge, no comparisons about the delay-utility performance between using the first-order part and the entire expression and between using the real Lyapunov function (L_2) and L_1 are given in the previous works of QLA. There are several works focusing on reducing the backlog of OQLA, e.g. [9][10].

We now summarize the main contributions of this paper in the following. 1) We prove that either using the entire Lyapunov drift expression instead of the first-order part of upper bound of Lyapunov drift of L_1 or using the real Lyapunov function instead of L_1 doesn't improve utility and delay performance. 2) We demonstrate and prove the utility and delay performance of QLA with Perturbed Data Queues (QLA-PDQ-P & QLA-PDQ-E) and QLA based on Entire expression of Lyapunov drift (QLA-E).

The rest of the paper is organized as follows: In Section II, we state our network model. After some preliminary information is given in Section III, we present our main results in Section IV, including a Lyapunov function and backlog-utility performance of variants of OQLA. Explanations of those results are given to show the rationality of OQLA and the relationship between attraction point and perturbation. In Section V, we prove results in Section IV. Section VI provides simulation results of QLA-PDQ-E, QLA-PDQ-P. We conclude in Section VII.

II. Network Model

In this section we specify the network model we use, which is also widely used in other works about QLA.

A. *Network States & Actions*

The network consists of r (r $\in \mathbb{Z}^+$ and is finite) queues. There are M network states forming the state set $\mathcal{S}$. Each state is denoted as $\mathrm{s_i}$, indicating the current network parameters. The network operates in slotted time.

Denote network state at time t as $s(t)$. We assume that $s(t)$ is stationary ergodic processes with finite state space and evolves according to some general probability law, under which there exists a steady state distribution of $s(t)$. Let p_{s_i} denote its steady state probability of being in state s_i, i.e,$p_{s_i} = \Pr\{s(t) = s_i\}$. At each timeslot t when the state $s(t) = s_i$, the network controller chooses an action $\alpha(t)$ from a set $\mathcal{A}^{s_i}$, i.e. $\alpha(t) = \alpha^{s_i}$ for some $\alpha^{s_i} \in \mathcal{A}^{s_i}$. The set $\mathcal{A}^{s_i}$ is called the feasible action set for network state s_i and is assumed to be time-invariant and compact for all $s_i \in \mathcal{S}$. Denote the action vector $< \alpha^{s_1}, \dots, \alpha^{s_M} >$ as $\boldsymbol{\alpha}$.

B. *Queues*

Let $\mathrm{Q(t)} =< \mathrm{Q_1(t)}, \dots, \mathrm{Q_r(t)} >$ denote the data queue backlog vector process of the network, where $\mathrm{Q_i(t)}$ is non-negative. Each queue is updated in the following way.$\mathrm{Q_i(t+1)} = \max\{\mathrm{Q_i(t)} - \mathrm{b_i(t)},0\} + \mathrm{A_i(t)}$. Queue i is mean rate stable (shorted for "stable" hereafter) means:$\lim_{t\to\infty} \frac{\mathbb{E}\{|Q_i(t)|\}}{t} = 0$[2]. The network is stable if all data queues are stable. Virtual queues can be used to represent time-averaged constraints. Simulations in Section VI consider a problem with time-averaged constraints to show our analysis and conclusions still hold.

Lyapunov function used in OQLA is defined as $L_1(t) = \frac{1}{2}\{\sum_{i=1}^{r} Q_i\,(t)^2\}$. Lyapunov drift of $L_1(t)$ is $D_1(t) = L_1(t+1) - L_1(t)$. The first-order part of upper bound of $D_1(t)$ used in OQLA is denoted as Δ_1.

Define Lyapunov function $L'(t)$ as $L'(t) = L'(t, \mathbf{C}) = \frac{1}{2}\{\sum_{i=1}^{r}(\,Q_i(t) - C_i)^2\}$, where $\mathbf{C} =< C_1(t), \dots, C_r(t) >$ denotes the perturbation of data queues. Lyapunov drift of the$L'(t)$ is $D'(t, \mathbf{C}) = L'(t+1) - L'(t)$. The first-order part of upper bound of $D'(t)$ is denoted as $\Delta'_{\mathbf{C}}$. Note that $L_1(t) = L'(t, \mathbf{0})$. Define Lyapunov function $L_2(t)$ as $L_2(t) = L'(t, \boldsymbol{\gamma}^*(V)_{\mathrm{OQLA}})$, Lyapunov drift of which is denoted as $D_2(t)$. The first-order part of upper bound of $D_2(t)$ is denoted as Δ_2. Expressions of $D_1(t)$, Δ_1, $D_2(t)$, Δ_2, $D'(t)$ and Δ' are shown and proved in [11].

C. *Stochastic Optimization Problem*

We consider a stochastic optimization problem with utility maximization. A utility function f is a function of network parameters, such as total throughput or energy-cost. Define time-average expectation of f(t) over the first T timeslots as $\bar{\mathrm{f}} = \frac{1}{\mathrm{T}}\sum_{\tau=1}^{\mathrm{T}} \mathbb{E}\,\{\mathrm{f}(\tau)\}$.

A network controller is designed to solve this problem, which operates a network with the goal of minimizing $\limsup_{t\to\infty}\overline{f}$, subject to the queue stability and additional time-average constraints. The case maximizing f can be treated the same way by letting $f' = -f$. We assume the network controller can observe $s(t)$ at the beginning of every timeslot t, but the p_{s_i} probabilities are not necessarily known. Thus f is can be regarded as a function of $s(t)$ and $\alpha(t)$, i.e. $f(t) = f(s(t), \alpha(t))$. Define f^{opt} as the maximum value of $\limsup_{t\to\infty}\overline{f}$ over all control policies that satisfies the stability and time-average constraints.

This problem is solved by OQLA in the way that each timeslot the network controller chooses the action that greedily minimizes $\Delta_1 + Vf$, where V is a positive constant $V \geq 1$ [2].

OQLA: $\alpha(t) = \mathrm{argmin}_{\alpha(t)\in\mathcal{A}^{s(t)}}\{\Delta_1(t) + Vf(t)\}$

It has been proved that OQLA stabilizes the network and achieves maximum utility asymptotically [2].

The following four variants of OQLA are analyzed in this paper.

- QLA-P: QLA using first-order Part of upper bound of Lyapunov drift of L_2 with a parameter V:
  $\alpha(t) = \mathrm{argmin}_{\alpha(t)\in\mathcal{A}^{s(t)}}\{\Delta_2(t) + Vf(t)\}$
- QLA-E: QLA using Entire expression of Lyapunov drift of L_1 with a parameter V:
  $\alpha(t) = \mathrm{argmin}_{\alpha(t)\in\mathcal{A}^{s(t)}}\{D_1(t) + Vf(t)\}$
- QLA-PDQ-P: QLA with Perturbed Data Queue using first-order Part of upper bound of Lyapunov drift of L' with parameters V and $\mathbf{C}$:
  $\alpha(t) = \mathrm{argmin}_{\alpha(t)\in\mathcal{A}^{s(t)}}\{\Delta'(t, \mathbf{C}) + Vf(t)\}$
- QLA-PDQ-E: QLA with Perturbed Data Queue using Entire expression of Lyapunov drift of L' with parameters V and $\mathbf{C}$:
  $\alpha(t) = \mathrm{argmin}_{\alpha(t)\in\mathcal{A}^{s(t)}}\{D'(t, \mathbf{C}) + Vf(t)\}$

Note that OQLA is equivalent to QLA-PDQ-P when $\mathrm{C} = 0$, QLA-E is equivalent to QLA-PDQ-E when $\mathrm{C} = 0$, QLA-P is equivalent to QLA-PDQ-P when $\mathrm{C} = \gamma^*(\mathrm{V})_{\mathrm{OQLA}}$.

D. *An Example of the Model*

Here we provide an example to illustrate our model, which will be used in Section VI. There are 3 nodes in the network. Each node can communicate with another. Thus there are 6 queues in the network. Denote the queue from node i to node j as $(\mathrm{i,j})$ $(\mathrm{i} \neq \mathrm{j})$, backlog of which is denoted as $\mathrm{Q_{ij}(t)}$. Possible data flows of queues are shown in Table [t_1].

TABLE I. Queues and Data Flows in a Network of 3 Nodes

Node No.	Queue No.	Possible Input		Possible Output	
N1	Q(1,2)	E[a]	Q(3,2)	N2	Q(3,2)
	Q(1,3)	E	Q(2,3)	N3	Q(2,3)
N2	Q(2,1)	E	Q(3,1)	N1	Q(3,1)
	Q(2,3)	E	Q(1,3)	N3	Q(1,3)
N3	Q(3,1)	E	Q(2,1)	N1	Q(2,1)
	Q(3,2)	E	Q(1,2)	N2	Q(1,2)

a. "E" is short for exogenous arrivals from outside the network

Network state consists of link state l_{ij} of from node i to node j, where $l_{ij} \in \{l_k, k = 1,2,3,4\}$. $l_k (k = 1,2,3,4)$ denote link status Good, Common, Bad and Disconnected respectively. Define $\xi_{ij} \in \{3,2,1,0\}$, where $\xi_{ij} = 3,2,1,0$ if $l_{ij} = l_1, l_2, l_3, l_4$ respectively. Link state is i.i.d. and $l_{ij} = l_k (k = 1,2,3,4)$ with equal probabilities. There are totally 4^6 network states.

Exogenous arrival into queue (i,j) from outside the network is denoted as a_{ij}. Maximum exogenous arrival to a queue is $A_{\max} = 6$, which means a_{ij} satisfies the constraint $0 \le a_{ij} \le A_{\max}$. Service allocated from queue (i,j) to node k is denoted as b_{ijk}. Power-service function is defined as $b_{ijk} = \ln\{1 + \xi_{ik} p_{ijk}\}$. Packets from queue (i,j) can only be transmitted to either its destination node j, or queue (k,j) of the other node k, which means $p_{iji} = 0$. Maximum power allocated to a queue is $P_{\max} = 6$, which means p_{ijk} satisfies $0 \le p_{ijk} \le P_{\max}$. Define the power out of node i as $p_i^N(t) = \sum_{j \ne i, k \ne i} p_{ijk}\,(t)$. Time-average power out of any node should be lower than P_{av}, i.e. $\text{limsup}_{t \to \infty} \frac{1}{t} \sum_1^t p_i^N\,(t) \le P_{\text{av}}$. Thus the corresponding virtual queue X_i updates according to $X_i(t+1) = \max\{X_i(t) + p_i^N(t) - P_{\text{av}}, 0\}$. There are totally 3 virtual queues. Utility function is defined as $f(t) = \sum_{i,j} \ln(1 + a_{ij})$ to represent total throughput of the network.

Each timeslot, according to maximum constraints of a_{ij}'s and p_{ijk}'s and time-average constraints of p_i^N network controller decides the amount of packets into each queue and decides the power allocated to each queue, i.e. network controller decides the values of a_{ij}'s and p_{ijk}'s.

III. Preliminary

A. Definitions

Let $\|\mathrm{Y}\|$ denote the inner product of Y, i.e. $\|\mathrm{Y}\| = \mathrm{Y} \bullet \mathrm{Y}^{\mathrm{T}}$.

Definition 1 (Attraction Point). Define the attraction point of a stochastic process Y(t) as follows.

Y^* is the Attraction Point of a process Y(t) with parameters ν, D and η if: There exist $\nu > 0$, $\mathrm{D} > \eta > 0$, such that $\mathrm{D} = \mathrm{D}(\nu)$, $\eta = \eta(\nu)$, and whenever $\|\mathrm{Y}(t) - \mathrm{Y}^*\| \ge \mathrm{D}$, we have $\mathbb{E}\{\|\mathrm{Y}(t + \mathrm{T}_\nu) - \mathrm{Y}^*\| - \|\mathrm{Y}(t) - \mathrm{Y}^*\| | \mathrm{Y}(t)\} \le -\eta$.

Definition 2 (Locally Polyhedral). Define locally polyhedral property same as in [9].

Definition 3. Define functions $\mathrm{l}_{\mathrm{k},\mathrm{s_i}}, \mathrm{g}_{\mathrm{k},\mathrm{s_i}}, \mathrm{l_k}, \mathrm{g_k} (\mathrm{k} = 0, \dots, 4)$ as follows.

$$\mathrm{l}_{0,\mathrm{s_i}}(\mathrm{Q}, \alpha^{\mathrm{s_i}}, \mathrm{V}) \quad = \{\mathrm{Vf}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}) + \sum_{\mathrm{i}=1}^{\mathrm{r}} \mathrm{Q_i}\,(\mathrm{A_i}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}) - \mathrm{b_i}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}))\}$$

$$\mathrm{l}_{1,\mathrm{s_i}}(\mathrm{Q}, \alpha^{\mathrm{s_i}}, \mathrm{V}) \quad = \{\mathrm{Vf}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}) + \sum_{\mathrm{i}=1}^{\mathrm{r}} (\,\mathrm{Q_i} - \gamma^*(\mathrm{V})_{\mathrm{OQLA}}) \times$$

$$(\mathrm{A_i}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}) - \mathrm{b_i}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}))\}$$

$$\mathrm{l}_{2,\mathrm{s_i}}(\mathrm{Q}, \alpha^{\mathrm{s_i}}, \mathrm{V}) \quad = \{\mathrm{Vf}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}) + \sum_{\mathrm{i}=1}^{\mathrm{r}} (\frac{1}{2}(\mathrm{A_i}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}) - \tilde{\mathrm{b}}_{\mathrm{i}}\,(\mathrm{s_i}, \alpha^{\mathrm{s_i}}))^2 +$$

$$\mathrm{Q_i}(\mathrm{A_i}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}) - \tilde{\mathrm{b}}_{\mathrm{i}}\,(\mathrm{s_i}, \alpha^{\mathrm{s_i}})))\}$$

$$\mathrm{l}_{3,\mathrm{s_i}}(\mathrm{Q}, \alpha^{\mathrm{s_i}}, \mathrm{V}, \mathrm{C}) \quad = \{\mathrm{Vf}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}) + \sum_{\mathrm{i}=1}^{\mathrm{r}} (\,\mathrm{Q_i} - \mathrm{C_i})(\mathrm{A_i}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}) - \mathrm{b_i}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}))\}$$

$$\mathrm{l}_{4,\mathrm{s_i}}(\mathrm{Q}, \alpha^{\mathrm{s_i}}, \mathrm{V}, \mathrm{C}) \quad = \{\mathrm{Vf}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}) + \sum_{\mathrm{i}=1}^{\mathrm{r}} (\frac{1}{2}(\mathrm{A_i}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}) - \tilde{\mathrm{b}}_{\mathrm{i}}\,(\mathrm{s_i}, \alpha^{\mathrm{s_i}}))^2 +$$

$$(\mathrm{Q_i} - \mathrm{C_i})(\mathrm{A_i}(\mathrm{s_i}, \alpha^{\mathrm{s_i}}) - \tilde{\mathrm{b}}_{\mathrm{i}}\,(\mathrm{s_i}, \alpha^{\mathrm{s_i}})))\}$$

Define $\mathrm{g}_{\mathrm{k},\mathrm{s_i}}$ as $\inf_{\alpha^{\mathrm{s_i}} \in \mathcal{A}^{\mathrm{s_i}}} \mathrm{l}_{\mathrm{k},\mathrm{s_i}}(\mathrm{Q}, \alpha^{\mathrm{s_i}})$. Define $\mathrm{g_k}(\mathrm{Q})$ as $\sum_{\mathrm{s_i}} \mathrm{p}_{\mathrm{s_i}}\, \mathrm{g}_{\mathrm{k},\mathrm{s_i}}$. Define $\mathrm{l_k}(\mathrm{Q}, \alpha)$ as $\sum_{\mathrm{s_i}} \mathrm{p}_{\mathrm{s_i}}\, \mathrm{l}_{\mathrm{k},\mathrm{s_i}}(\mathrm{Q}, \alpha^{\mathrm{s_i}})$. Note that $\mathrm{g_k} = \inf_{\alpha} \mathrm{l_k}(\mathrm{Q}, \alpha^{\mathrm{s_i}})$.

B. Assumptions

We list the assumptions used hereafter, which are as same as the ones in [8]. These assumptions hold in many network utility optimization problems and aren't so rigorous as they appear. Explanations can be found in [11].

Assumption 1. Local maximum point of $g_i (i = 0, \dots, 4)$ is unique on $\mathbb{R}^r$, denoted as $\boldsymbol{\gamma}_i^*(V)$.

Assumption 2. $g_i (i = 0, \dots, 4)$ is locally polyhedral at its maximum point.

Assumption 3. ϵ-slackness[9] holds for the network.

Assumptions 1 and 2 are about the property of utility function, while Assumption 3 is about the network.

C. Lemmas

Before moving further we introduce the following three lemmas, proofs of which are omitted for brevity and can be found in [2][3][8].

Lemma 1. $\boldsymbol{\gamma}^* \ge \mathbf{0}$ is the unique attraction point of $\mathbf{Q}(t)$ if:

1) A function $g(\mathbf{Q})$ is locally polyhedral at $\boldsymbol{\gamma}^*$.
2) For all s_i, there exists a positive constant C satisfying

$$\| \mathbf{Q}^{s_i}(t+1) - \boldsymbol{\gamma}^* \|^2 - \| \mathbf{Q}(t) - \boldsymbol{\gamma}^* \|^2 \le C - 2(g_{s_i}(\boldsymbol{\gamma}^*) - g_{s_i}(\mathbf{Q}(t)))$$

Lemma 2. For network under OQLA with a parameter V, for any point $\mathbf{Q}$, we have

$$g_0(\mathbf{Q}, V) \le g_0(\boldsymbol{\gamma}_0^*(V), V) \le V f^{\text{opt}}$$

IV. Main Results

We summarize analysis results here.

Denote the Euclid ball centered at $\mathbf{Q}_0$ with radius r as $\mathbf{B}(\mathbf{Q}_0, r)$.

Theorem 1. When the network state is i.i.d., in the sense of conditional expectation, one Lyapunov function of the network under OQLA is

$$L_3(\mathbf{Q}) = \begin{cases} L_2 & \mathbf{Q} \notin \mathbf{B}(\boldsymbol{\gamma}_0^*(V), D_0) \\ 0 & \mathbf{Q} \in \mathbf{B}(\boldsymbol{\gamma}_0^*(V), D_0) \end{cases}$$

Note that L_2 equals L_3 in most cases thus L_2 is used in QLA-P instead of L_3.

Theorem 2. For the network under QLA-P (QLA using first-order Part of upper bound of Lyapunov drift of L_2) with a parameter V, the following properties hold. The attraction point of queue backlog is $2\boldsymbol{\gamma}_0^*(V)$ Utility function satisfies $\text{limsup}_{t \to \infty} \overline{f} \le f^{\text{opt}} + B_1/V$ where B_1 is a positive constant.

Theorem 2 shows that using L_2 instead of L_1 when designing QLA, queue backlog is doubled while utility performance remains $\mathcal{O}(1/V)$ compared to OQLA. Thus using the Lyapunov function of the network under OQLA does not help the queue backlog and utility performance, only causing backlog to be larger.

Theorem 3. For the network under QLA-E (QLA using Entire expression of Lyapunov drift of L_1) with a parameter V, the following properties hold. Queue backlog is attracted by $\boldsymbol{\gamma}_2^*(V)$. $\boldsymbol{\gamma}_2^*$ approximately equals $\boldsymbol{\gamma}_0^*$ when V is large enough. Utility function satisfies $\limsup_{t\to\infty}\overline{f} \le f^{\text{opt}} + B_2/V$, where B_2 is a positive constant.

Theorem 3 shows that using the entire Lyapunov drift expression instead of the first-order part of the upper bound of Lyapunov drift of L_1 when designing QLA, queue backlog and utility performance is not improved compared to OQLA. However, when using the entire expression, some problems with variables coupled loosely which can be solved in a distributed manner become problems with variables coupled tightly which can only be solved in a centralized manner. Thus using the entire expression increases the complexity. Therefore using the first-order part when designing QLA (such as in [2]) is reasonable.

Theorem 4. For the network under QLA-PDQ-P (QLA with Perturbed Data Queue using first-order Part of upper bound of Lyapunov drift of L') with parameters V and $\mathbf{C}$, the following properties hold. If $\boldsymbol{\gamma}_0^*(V) \ge -\mathbf{C}$, the attraction point of queue backlog $\boldsymbol{\gamma}_3^*$ equals $\boldsymbol{\gamma}_0^*(V) + \mathbf{C}$. Utility function satisfies $\limsup_{t\to\infty}\overline{f} \le f^{\text{opt}} + \frac{B_3}{V}$, where B_3 is a positive constant.

Theorem 4 shows that using a positive $\mathbf{C}$ increases backlog while using a negative $\mathbf{C}$ decreases backlog. This idea is used when designing QLA-VPDQ.

Theorem 5. For the network under QLA-PDQ-E (QLA with Perturbed Data Queue using Entire expression of Lyapunov drift of L') with parameters V and $\mathbf{C}$, the following properties hold. The attraction point of queue backlog $\boldsymbol{\gamma}_4^*$ equals $\boldsymbol{\gamma}_2^*(V) + \mathbf{C}$, Utility function satisfies $\limsup_{t\to\infty}\overline{f} \le f^{\text{opt}} + \frac{B_4}{V}$, where B_4 is a positive constant.

Theorem 5 shows that using the entire expression doesn't help to enhance queue backlog and utility performance even for QLA with perturbed data queue.

V. PROOFS

A. *Proof of Theorem 1*

It can be seen from the definition of L_3 that $L_3 \ge 0$ and $L_3 = 0$ only when $Q \in B(\gamma_0^*(V), D_0)$. Thus we have from Lemma 1 that $\mathbb{E}\{\|Y(t+1) - Y^*\| - \|Y(t) - Y^*\| | Y(t)\} < 0$ when $Q \notin B(\gamma_0^*(V), D_0)$.

Thus L_3 is a Lyapunov function of the network in the sense of conditional expectation with the stability point $\mathbf{B}(\boldsymbol{\gamma}_0^*(V), D_0)$.

B. *Proof of Theorem 3*

Lemma 3. For the network under QLA-E, the following two equations hold.

For any Q_1 and Q_2, we have

$$g_2(Q_1) \le g_2(Q_2) + (Q_1 - Q_2) \bullet (A(t) - \tilde{b}(t)) \quad (1)$$

For any Q, we have

$$\begin{aligned}\| Q^{s_i}(t+1) \quad - \gamma_2^* \|^2 - \| Q(t) - \gamma_2^* \|^2 \le \\ C_2 - 2(\gamma_2^* - Q(t)) \bullet (A(t) - \tilde{b}(t))\end{aligned} \quad (2)$$

Proof of Lemma 3 can be found in [11].

1)Property of Queue Backlog

Attraction Point Property

From (1) and (2) we have for any Q

$$\begin{aligned}\| Q^{s_i}(t+1) - \gamma_2^* \|^2 \quad - \| Q(t) - \gamma_2^* \|^2 \le \\ C_2 - 2(g_{2,s_i}(\gamma_2^*) - g_{2,s_i}(Q(t))\end{aligned}$$

Using Lemma 1 and noting that g_2 is polyhedral it can be concluded that γ_2^* is the unique attraction point of queue backlog Q(t).

Linear Property of Attraction Point

From the expression of g_2 we see that,

$$\begin{aligned} g_2(Q)/V \quad = \inf_{\alpha} \sum_{s_i} p_{s_i} \{f(s_i, \alpha^{s_i}) + \\ \sum_{i=1}^{r} (\frac{1}{2V}(A_i(s_i, \alpha^{s_i}) - \tilde{b}_i(s_i, \alpha^{s_i}))^2 + \\ Q'_i(A_i(s_i, \alpha^{s_i}) - \tilde{b}_i(s_i, \alpha^{s_i})))\} \end{aligned}$$

where $Q'_i = \frac{Q_i}{V}$. When V is large enough and $\frac{1}{2V}(A_i(s_i, \alpha^{s_i}) - \tilde{b}_i(s_i, \alpha^{s_i}))^2$ is small enough to be ignored, the right hand side is $g_0(Q', 1)$ which maximized at $\gamma_0^*(1)$. Therefore we have $\gamma_2^*(V) \approx V\gamma_0^*(1) = \gamma_0^*(V)$.

2)Property of Utility Function

By [11], we have

$$\begin{aligned} &\mathbb{E}\{D_1 + Vf | Q(t)\} \\ &\le \frac{1}{2}\mathbb{E}\{A_i(t)^2 + b_i(t)^2 - 2A_i(t)\tilde{b}_i(t) + \\ &2Q_i(t)(A_i(t) - b_i(t)) | Q(t)\} \\ &\le B'^2 + \mathbb{E}\{Q_i(t)(A_i(t) - b_i(t)) | Q(t)\} \end{aligned}$$

Because QLA-E greedily minimizes $D_1 + Vf$, we have

$$\begin{aligned} &\mathbb{E}\{D_1 + Vf | Q(t)\} \\ &\le B'^2 + \mathbb{E}\{Q_i(t)(A_i(t)^{ALT} - b_i(t)^{ALT}) | Q(t)\} \end{aligned}$$

where ALT represents any other alternate policy.

Now using OQLA as ALT, using Lemma 2, we have

$$\begin{aligned} &\mathbb{E}\{D_1 + Vf | Q(t)\} \le B'^2 + g_0(Q(t), V) \\ &\le B'^2 + Vf^{opt} \end{aligned}$$

Taking expectations over Q(t) and summing the above over $t = 1, \ldots, T-1$, we have:

$$\mathbb{E}\{L_1(T) - L_1(1)\} + V\sum_{\tau=1} T f(\tau) \le TB'^2 + VTf^{opt}$$

Rearranging the terms, using the facts that $L(t) \ge 0$ and $L(0) = 0$, dividing both sides by VT, and taking the limsup as $T \to \infty$, we get:

$$\limsup_{T\to\infty} \bar{f} \le f^{opt} + \frac{B'^2}{V}$$

Proof completes by letting $B_2 = B'^2$.

C. *Proof of Theorem 4*

Lemma 4. For the network under QLA-PDQ-P, the following two equations hold.

For any Q_1 and Q_2, we have

$$g_3(Q_1) \le g_3(Q_2) + (Q_1 - Q_2) \bullet (A(t) - b(t)) \quad (3)$$

For any Q, we have

$$\| Q^{s_i}(t+1) \quad -\gamma_3^* \|^2 - \| Q(t) - \gamma_3^* \|^2 \le C_3 - 2(\gamma_3^* - Q(t)) \bullet (A(t) - b(t)) \quad (4)$$

Proof of Lemma 4 can be found in [11].

3) Property of Queue Backlog

From (3) and (4) we have for any Q

$$\| Q^{s_i}(t+1) - \gamma_3^* \|^2 \quad - \| Q(t) - \gamma_3^* \|^2 \le C_2 - 2(g_{3,s_i}(\gamma_3^*) - g_{3,s_i}(Q(t))$$

Using Lemma 1 and noting that g_3 is polyhedral it can be concluded that γ_3^* is the unique attraction point of queue backlog $Q(t)$. Noting that $g_1(Q) = g_3(Q - C)$. Thus $\gamma_3^* = \gamma_0^* + C$ if $\gamma_0^* \ge -C$.

4) Property of Utility Function

By [11], we have

$$\mathbb{E}\{\Delta' + Vf|Q(t)\}$$
$$\le \frac{1}{2}\mathbb{E}\{A_i(t)^2 + b_i(t)^2 - 2A_i(t)\,\tilde{b}_i\,(t) +$$
$$2(Q_i(t) - C_i)(A_i(t) - b_i(t))|Q(t)\}$$
$$\le B'^2 + \mathbb{E}\{(Q_i(t) - C_i)(A_i(t) - b_i(t))|Q(t)\}$$

Because QLA-PDQ-P greedily minimizes $\Delta' + Vf$, we have

$$\mathbb{E}\{\Delta' + Vf|Q(t)\}$$
$$\le B'^2 + \mathbb{E}\{(Q_i(t) - C_i)(A_i(t)^{ALT} - b_i(t)^{ALT})|Q(t)\}$$

where ALT denotes represents any other alternate policy.

Now using OQLA as ALT, using Lemma 2, we have

$$\mathbb{E}\{\Delta' + Vf|Q(t)\} \le B'^2 + g_0(Q(t) - C)$$
$$\le B'^2 + Vf^{opt}$$

Following the same line as in Section V-B, we have

$$\limsup_{T\to\infty} \bar{f} \le f^{opt} + \frac{B'^2}{V}$$

Proof completes by letting $B_3 = B'^2$.

D. Proof of Theorem 2

Using Theorem 4 while letting $C = \gamma_0^*$ completes the proof.

E. Proof of Theorem 5

1)Property of Queue Backlog

Following the same line as in Section V-C, the relationship between γ_2^* and γ_4^* can be obtained.

2)Property of Utility Function

Similar to the one in Section V-C. The difference is that QLA-PDQ-E greedily minimizes $D' + Vf$.

VI. Simulation

In this section we provide simulation results for the QLA-PDQ-P, QLA-PDQ-E and QLA-VPDQ on the network model in Section II-D. QLA-P and QLA-E are omitted here because they are specific form of QLA-PDQ-P and QLA-PDQ-E respectively. We simulate QLA-PDQ-P and QLA-PDQ-E with $V_i = 10 + 100i (i = 0, \dots, 9)$ and $C_j = -500 + 100j (j = 0, \dots, 8)$, where bold symbol means a vector with all components equaling the same constant. Precision of p_{ijk} and a_{ij} is set to 0.01. We run each case for 10^4 timeslots under both algorithms. Under each value of V and C, average queue backlog and utility function are obtained by using the final 5000 timeslots when the network is in the steady-state (Fig. 1 and 2).

Linear relationship between attraction point and $\mathbf{C}$ mentioned in Theorem 4 and 5 are shown in Fig. 1(b) and 2(b). Linear relationship between attraction point and V mentioned in Theorem 2 and 3 are shown in Fig. 1(a) and 2(a). From Fig. 1(b) and 2(b), Fig. 1(a) and 2(a) it can be seen that $\boldsymbol{\gamma}_2^*$ approximately equals $\boldsymbol{\gamma}_0^*$ as mentioned in Theorem 3.

It can be seen from Fig. 1(d) and 2(d) that f decreases dramatically if $\mathbf{C}$ decreases when $\mathbf{C} \le -\boldsymbol{\gamma}_0^*$. However, theoretical relationship between $\mathbf{C}$ and f when $\mathbf{C} \le -\boldsymbol{\gamma}_0^*$ remains an open question. This question may relate to the property of g_0. However, we can see from Fig. 1(c) and 2(c) that $f \to f^{opt}$ as $V \to \infty$ for all values of $\mathbf{C}$, as mentioned in Theorem 4 and 5.

VII. Conclusion

We have investigated several variants of OQLA in this paper. First, rationality of OQLA is proved for using the first-order part of upper bound of drift of a function. Although the entire expression of drift is not used in OQLA, backlog and utility of OQLA performance is the same. Although the Lyapunov function of the network is not used in OQLA, backlog of OQLA halved and utility performance is the same. Therefore it is of no need to use either the entire expression or the Lyapunov function of the network. Second, linear relationship between perturbation in data queues and attraction point of the backlog is found.

Acknowledgment

The work described in this paper was supported by the National Natural Science Foundation of China (No. 61471023).

References

[1] L. Georgiadis, M. J. Neely, and L. Tassiulas, Resource allocation and cross-layer control in wireless networks. Now Publishers Inc, 2006.

[2] M. J. Neely, Stochastic network optimization with application to communication and queueing systems. Morgan &Claypool Publishers, 2010, vol. 3, no. 1.

[3] ——, “Energy optimal control for time-varying wireless networks,” IEEE Transactions on Information Theory, vol. 52, no. 7, pp. 2915–2934, 2006.

[4] R. Urgaonkar and M. J. Neely, “Optimal routing with mutual information accumulation in wireless networks,” in ConferenceRecord of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR),. IEEE, 2011, pp. 1602–1609.

[5] L. Huang and M. J. Neely, “Utility optimal scheduling in energy harvesting networks,” in Proceedings of the Twelfth ACM International Symposium on Mobile Ad Hoc Networking and Computing. ACM, 2011, p. 21.

[6] ——, “Utility optimal scheduling in processing networks,” Performance Evaluation, vol. 68, no. 11, pp. 1002–1021, 2011.

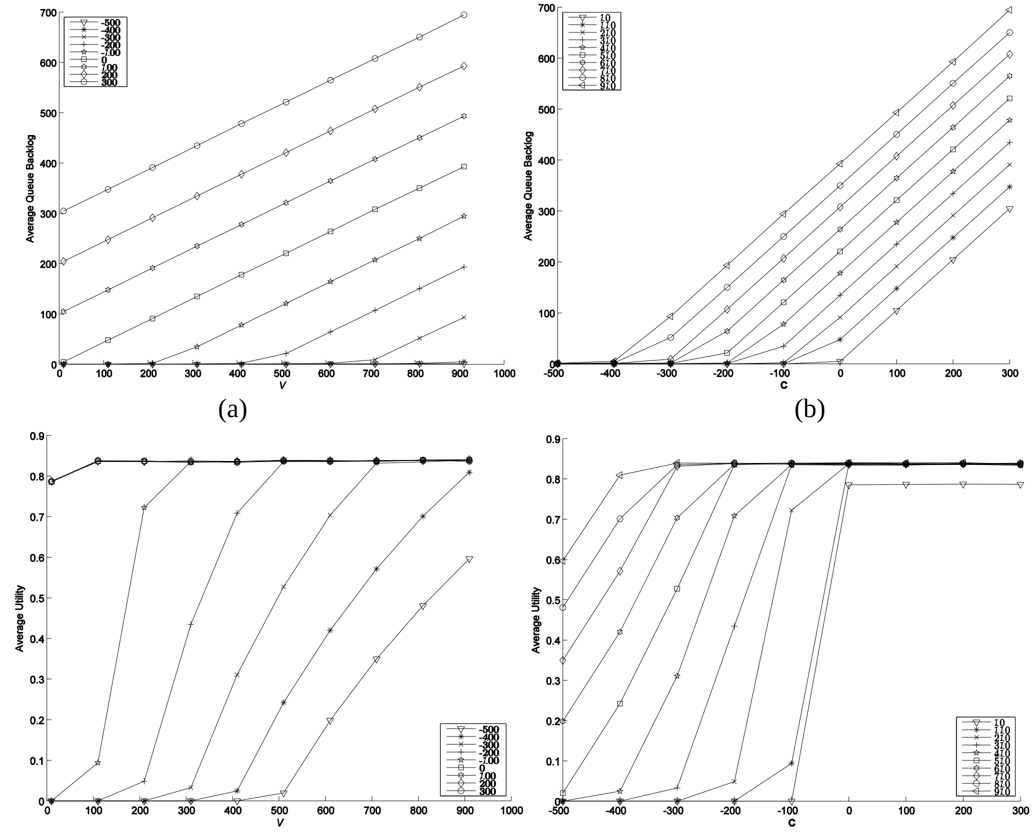
[7] M. J. Neely, “Stock market trading via stochastic network optimization,” in 49th IEEE Conference on Decision and Control (CDC). IEEE, 2010, pp. 2777–2784.

[8] L. Huang and M. J. Neely, "Delay reduction via lagrange multipliers in stochastic network optimization," IEEE Transactions on Automatic Control, vol. 56, no. 4, pp. 842–857, 2011.

[9] L. Huang, "Deterministic mathematical optimization in stochastic network control," Ph.D. dissertation, UNIVERSITY OF SOUTHERN CALIFORNIA, 2011.

[10] M. J. Neely and R. Urgaonkar, "Opportunism, backpressure, and stochastic optimization with the wireless broadcast advantage," in 42nd Asilomar Conference on Signals, Systems and Computers. IEEE, 2008, pp. 2152–2158.

[11] LI Hu and etc.. "The detailed version of 'Equivalence on Quadratic Lyapunov Function Based Algorithms in Stochastic Networks'", https://figshare.com/s/4d4d27f38d5d07166c05, available since March 2017.



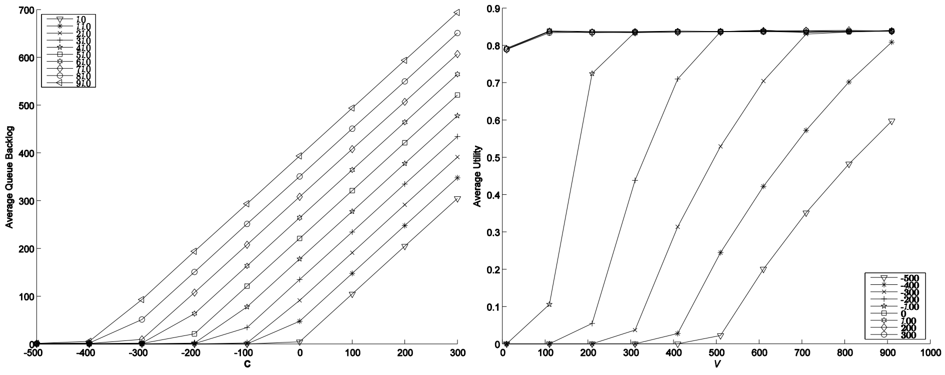
(a)

(b)

(c)

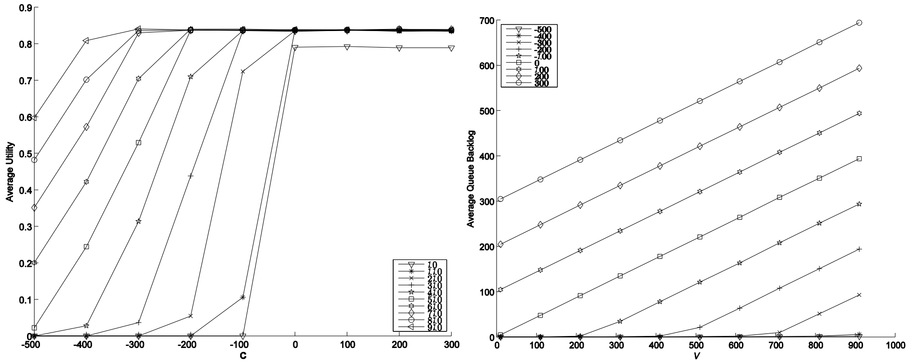
(d)

Figure 1. Simulation Results for QLA-PDQ-P: (a) Average Queue Backlog vs. V; (b) Average Queue Backlog vs. $\boldsymbol{C}$; (c) Average Utility vs. V; (d) Average Utility vs. $\boldsymbol{C}$



(a)

(b)



(c) (d)

Figure 2. Simulation Results for QLA-PDQ-E: (a) Average Queue Backlog vs. V ; (b) Average Queue Backlog vs. C; (c) Average Utility vs. V ; (d) Average Utility vs. C.

A Remote-Attestation-Based Extended Hash Algorithm for Privacy Protection

Yongxiong Zhang
Department of Economics and Trade
Guangzhou College of Technology and Business,
GCTB
Guangzhou, China
E-mail: csyxzhang@qq.com

Yucong You
Department of Economics and Trade
Guangzhou College of Technology and Business,
GCTB
Guangzhou, China
E-mail: 61070262@qq.com

Liangming Wang *
School of Software Engineering
South China University of Technology, SCUT
Guangzhou, China
E-mail: lmwang@scut.edu.cn
*The corresponding author

Luxia Yi
Department of Economics and Trade
Guangzhou College of Technology and Business,
GCTB
Guangzhou, China
E-mail: cilu-5@qq.com

***Abstract*—Compared to other remote attestation methods, the binary-based approach is the most direct and complete one, but privacy protection has become an important problem. In this paper, we presented an Extended Hash Algorithm (EHA) for privacy protection based on remote attestation method. Based on the traditional Merkle Hash Tree, EHA altered the algorithm of node connection. The new algorithm could ensure the same result in any measure order. The security key is added when the node connection calculation is performed, which ensures the security of the value calculated by the Merkle node. By the final analysis, we can see that the remote attestation using EHA has better privacy protection and execution performance compared to other methods.**

Keywords-Trusted computing; Remote attestation; Privacy protection; Merkle hash tree; Extended hash algorithm

I. INTRODUCTION

With the extensive and widely use of Cloud Computing technology, more and more application services require completion of the interaction between multiple machines to achieve the tasks, how to ensure the machine's mutual-trustworthiness and security during the process has become a very important problem. The traditional security mechanisms are basically built upon the application level, which is difficult to ensure the machine's mutual-trustworthiness and security. Trusted Computing Group (Trusted Computing Group, referred to as TCG) [1] has proposed the trusted computing (referred to as TC) technology development, leading to more and more system security solutions based on trusted computing technology, of which the trusted root and trust chain delivery mechanism provides a basic environment for the solution. The remote attestation technology in trusted computing provides a solution to the trust of both parties. The standard remote attestation mechanism given by the Trusted Computing Specification can be divided into three steps: (1) integrity status measurement; (2) measurement results query communication; (3) integrity status attestation. Those three steps constitute, as a whole, a remote certification mechanism framework. The integrity state measure is mainly to collect information on the hardware and software stack integrity status of the platform so as to be verified, mainly through a large number of Hash operations, and the measurement process to generate specific results stored in the TPM PCR. The integrity measure functions as the basis of the entire telematics technology, being a reliable attestation of whether it is safe and effective.

In 2004, based on the research of Trusted Computing, the IBM Research Center presented the design of the Integrity Measurement Architecture (IMA) [2] based on the Trusted Platform Module TPM, which was designed on the Linux operating system to accomplish the task. When the system opens the file setting into the memory, the IMA code set in the system will evaluate the integrity of the file, then saves the measurement results in the metric list, meanwhile it extends the metric to the TPM chip. The integrity of the IMA definition is based on the simple metric load code and some system static data, moreover, the IMA inserts a large number of metric points when performing the integrity measure, and thus increases the inaccuracy of the metric and the redundancy of the metric. Both of the IMA and other customary binary method, conduct the operation of each file via the simple Hash connection during the process of measurement , the attestation process requires the entire the process log, introducing the privacy exposure , and because the attestation of the log requires the re-completion of the entire process, the performance tends to be inferior. As for the attribute-based remote attestation in [3,4], the proving party only needs to give the corresponding attribute declaration, which is according to the target attribute of the verifier, and does not need to expose the entity uniqueness mark to the verifier. At the same time, because the unique identity of the running entity in the system can be regarded as one of the attributes of the system, the attributes-based remote attestation improves the flexibility of the certification program on the protection of system privacy. However, there exists a problem to perfect the completeness of attributes-based remote attestation method. The other methods in [5,6,7] also cannot solve the completeness problem.

Based on a more direct and complete binary remote attestation method, this paper proposes an Extended Hashing Algorithm with a more efficient and privacy protection. Extended Hashing Algorithm will be used in the process of integrity measurement in the process of remote attestation.

The rest of this paper is organized as follows. Section II introduces relevant background knowledge. The third section introduces the Extended Hashing Algorithm based on Merkle Hash Tree. In the fourth section, we analyze the Extended Hashing Algorithm proposed in this paper. Section 5 summarizes this paper and prospects for the next step.

II. Background

A. Binary-based Remote Attestation

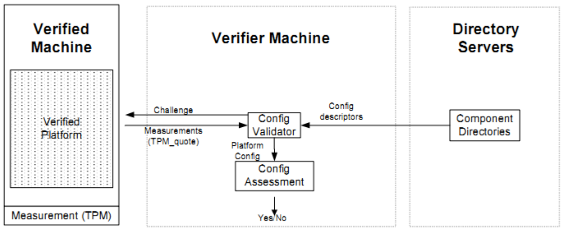


Figure 1. Framework of TCG’s binary-based remote attestation

As it is shown in Fig .1, in the Framework of binary remote authentication, the Verifier Machine and the Verified Machine communicate directly through the Config Validator module. The Config Validator module obtains the results of the binary measurement by the proving party through the TPM and reconstructs the configuration information of the platform. During the refactoring process, both of the report log information generated by the TPM metric process and the component configuration information provided by Direcory Servers are required to be used.

B. Merkle Hash Tree

Hash Tree, also known as the Merkle Hash Tree [8], was coined by Merkle as a method to establish the shared secret between the two entities by using a public key infrastructure in 1980. Merkle Hash Trees are now commonly used to protect the data blocks in memory [9,10,11]. It is shown by Figure 2 that a binary Merkle Hash Tree with four-leaf nodes. Apparently ,on the Merkle Hash Tree, the data object is created as a leaf node, and the tree's internal node is its sub node is the Hash Value concatenation. The root of the tree is named the “root hash”, which represents all data objects, because changes to arbitrary data objects tends to cause changes in the root Hash Value.

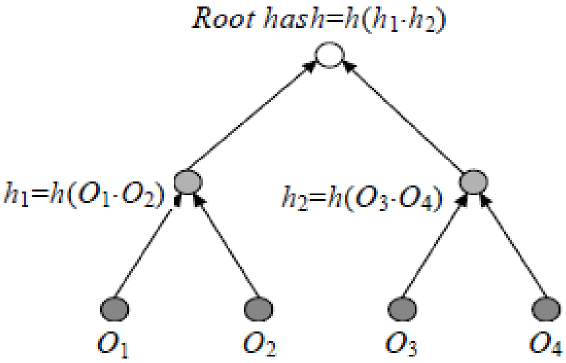


Figure 2. Merkle Hash Tree

To check the integrity of the leaf nodes, the following points are needed: (1) reading the contents of the node and its brother's nodes; (2) connecting their data; (3) calculating the Hash Value of the data after the connection; 4) Repeating the above steps to the Root Hash; (5) Checking whether the calculated result is consistent with the content stored in the Root Hash node.

To update the leaf node value, firstly check the integrity of the leaf node (process as mentioned above). If the leaf node is complete, then the followings are needed: (1) modifying the node value; (2) calculating and updating the value of its parent node; (3) repeating the above steps until the Root Hash is eventually updated.

III. A Merkle Hash Tree -Based Extended Hash Algorithm

As in [12], the calculation method of Hash Value of the inner node connection on the Merkle Hash Tree generates the h (o1.o2) <> h (o2.o1), so the different order of the leaf node data read during the process generated on the Merkle Hash Tree leads to different results.

The extended Merkle Hash Tree uses an alternative method when connecting joints between nodes in the generation process so as to ensure that the order of the nodes' data is the same.

A. Implementation of EHA Algorithm

EHA (A, B, K), A and B represent two node extended, k represent the security key.

Definition:

$N_L=\{0,1\}^{160}$, the leaf node on the tree, representing the 160-bit Hash Value of the file

$N_B=\{0,1\}^{168}$, the inner node in the tree, representing the 168-bit Hash result obtained by connecting the Hash by two other nodes

K, key

$K^*=\{0,1\}^{160}$, which represents the 160-bit Hash Value produced by the Hash of the key K

Algorithmic process:

Calculating the Hash of K to get the 160-bit Hash result K *

Calculating the number of bits in 1 in K *, denoted as CK

1) Connection between leaf nodes

NL1, NL2 represent , respectively , two leaf nodes,

*a) The NL1 and K * bits are XOR (XOR), and the result is recorded as S1*

*b) The NL2 and K * bitwise exclusive OR, the result is recorded as S2*

c) Calculate the number of bits in the NL1 and NL2 values of 1, denoted as C1, C2, respectively

d) S1 cycle left shift C1 bit, the result is recorded as CS1

e) S2 cycle left shift C2 bit, the result is recorded as CS2

f) The CS1 and CS2 are exclusive-OR, and the result is denoted by R

g) Calculate the number of bits in the value of R for 1, denoted as CR, CR is in the range of 0 to 160, so it can be represented by 8-bit binary

h) Connect the 160-bit R to the 8-bit CR to get 168 bits, denoted as Y

i) Move the Y cycle to the left to get the final result

2) Connection between the nodes

NB1, NB2 represent, respectively , two inner nodes

*a) NB1 cycle right shift CK bit, the result is recorded as NB1 **

*b) Take NB1 * before the first 160, recorded as NC1*

*c) NB2 cycle right shift CK bit, the result is recorded as NB2 **

*d) Take NB2 * before the 160, recorded as NC2*

e) The NC1 and NC2 are exclusive-OR, and the result is denoted by R

f) Calculate the number of bits of value 1 in R, denoted as CR

g) The 160-bit R and 8-bit CR connected to the results of 168, recorded as Y

h) Move the Y cycle to the left by the CK bit to get the final result

3) The connection between the leaf node and the inner node

NL, NB decibels represent leaf nodes and inner nodes

*a) NB cycle right shifts CK bit, the results recorded as NB **

*b) Take NB * the 160 downwards, recorded as NC*

c) Calculate the number of bits in the NL with a value of 1, denoted as C

*d) NL and K * bitwise exclusive OR, the result is recorded as S*

e) S cycle left shift C bit, the result is recorded as CS

f) The CS and NC bitwise exclusive OR, the result is recorded as R

g) Calculate the number of bits of value 1 in R, denoted as CR

h) Connect the 160-bit R to the 8-bit CR to get 168 bits, denoted as Y

i) Move the Y cycle to the left to get the final result

B. *Production Process Analysis of Extended Merkle Hash Tree*

In light of the definition of Extended Merkle Hash Tree, we can see that any extension of the Merkle Hash Tree , during the production process , is a number of times between the leaves of the connection between the Hash operation, a number of connections between the nodes of the Hash operation and a number of inner nodes and a leaf node, and the inner node is generated by the connection operation of two leaf nodes or one leaf node to another inner node.

By the description of the algorithm of the previous connection operation, it can be seen that the operation of step a , b , c , d of the connection operation between two inner nodes are the reverse operation of step 1 , 2of the operation of one leaf node and one inner node during the last two steps of the operation of, so in order to facilitate the analysis, we put the above three operations simplified process as follows , not affecting the results of the case:

- $\oplus$, XOR operation symbol
- <<<, loop left shift operation symbol

N1, N2 between the two leaves of the operation can be expressed as: $((N1 \oplus K*) <<< C1) \oplus ((N2 \oplus K*) <<< C2)$;

N1 and N2 two nodes within the operation of $N1 \oplus N2$;

The connection operation between the leaf node N1 and the inner node N2 $((N1 \oplus K*) <<< C1) \oplus N2$.

Assuming that the process of Extended Merkle Hash Tree is generated by the leaf node, we use the operator * to describe that the Extended Merkle Hash Tree that generates n leaf nodes is N1 * N2 ••• * Nn.

N1 * N2 is actually the connection between the two leaves, through the previous algorithm Definition: $N1 * N2 = ((N1 \oplus K*) <<< C1) \oplus ((N2 \oplus K*) <<< C2)$, Because the XOR operation is consistent with the exchange law, so $N1 * N2 = ((N1 \oplus K*) <<< C1) \oplus ((N2 \oplus K*) <<< C2) = ((N2 \oplus K*) <<< C2) \oplus ((N1 \oplus K*) <<< C1) = N2 * N1$.

N1 * N2 * N3 by N1 and N2 to do a connection between the leaves of the operation, and then the results and N3 do a leaf and the connection between the nodes in the operation.

By the previous definition, $N1 * N2 * N3 = ((N1 \oplus K*) <<< C1) \oplus ((N2 \oplus K*) <<< C2) * N3 = ((N1 \oplus K*) <<< (N2 \oplus K*) <<< C2) * ((N3 \oplus K*) <<< C3) = N1 * (N2 * N3)$.

$N1 * N2 = X1 \oplus X2$, $N1 * N2 * N3 = X1 \oplus X2 \oplus X3$, further result can be obtained, $N1 * N2 * ... * Nn = X1 \oplus X2 \oplus ... \oplus Xn$.

Thus, in the process of generating the Extended Merkle Hash tree, the result of the root node of the EMT is ultimately generated by the n leaf nodes, being consistent regardless of the order of the N_i.

IV. Analysis of the Advantages of Extended Hash Algorithm

A. *Privacy Protection Analysis*

The traditional TCG remote attestation mechanism extends the metric Hash Value of the metric into the PCR and reconstructs the PCR value when performing integrity verification. The advantage of this mechanism is that it is easy to construct an integrity trust chain. The drawback of it lies at it being necessary to understand the total integrity Hash of the metric and its extension to the PCR, so it is appropriate to measure the entire process from powering up the machine to the start of the operating system, but for the application, it does not trust mutually and generally do not have a strict execution order relationship, and the IMA metric-based metric verification mechanism still requires the application of the integrity of the Hash Value during the whole process aiming to extend the specified PCR to implement the measurement verification, which is the root cause of insufficient privacy protection.

Extended Hash Algorithm is used to remotely attest that the integrity of the program to be verified so as to calculate the integrity of the program, followed by a Merkle Hash Tree developed, and tree nodes preserve the integrity of the program Hash Value, and the tree in the non-leaf nodes are automatically generated by the Extended Hash Algorithm. When the remote attestation is conducted, there is only a node is obtained through the encrypted Hash Value.

Hence, it is clear that the use of Extended Hash Algorithm for remote attestation of privacy can be effectively protected.

B. *Analysis of Performance*

For the IMA architecture, a list of n nodes is required to perform this Extended Hash operation with a time performance of O (n). For the Extended Hash Algorithm based on the Merkle Hash Tree, the time performance is O (log2n). And this has proved that the implementation is of high efficiency.

V. Conclusion

This paper conducts an analysis on the existing problem of binary-based and attribute-based remote attestation method, because the attribute-based remote attestation method faces a complete problem, which is difficult to solve. To the binary-based remote attestation method, this paper focused the solution to the privacy protection problem. An Extended Hash Algorithm for privacy protection has been proposed. In the process of integrity measurement, the Extended Hash Algorithm uses the Merkle Hash Tree to store the binary Hash Value. During the process of generating the Merkle Hash Tree, compared to other algorithms, the Extended Hash Algorithm proposed in this paper is not affected by the extended order. Through the final security analysis, the algorithm achieves the desired effect.

In the next process, we will, apply the Extended Hash Algorithm proposed in this paper, to the remote attestation process, aiming to achieve the definition of the completion of the entire remote attestation.

Acknowledgment

This paper is supported by: key project of the Combination of “Production, Education and Research” supported by Guangdong province and Ministry of Education(project number: 2012B091000109); the second batch of core technology research project on strategic emerging industry of Guangdong Province(project number: 2012A010701005); the Fundamental Research Funds for the Central Universities(project number:2012ZM0051).

References

[1] “Trusted computing.” [Online]. Available: http://www. trustedcomputinggroup.org/

[2] R. Sailer, X. Zhang, T. Jaeger, and L. Van Doorn, “Design and implementation of a tcg-based integrity measurement architecture.” in USENIX Security Symposium, vol. 13, 2004, pp. 223–238.

[3] L. Chen, R. Landfermann, H. Lohr, M. Rohe, A.-R. Sadeghi, and ¨C. Stuble, “A protocol for property-based attestation,” in ¨ Proceedings of the first ACM workshop on Scalable trusted computing. ACM, 2006, pp. 7–16.

[4] Sadeghi A, Stüble C. Property-Based attestation for computing platforms: caring about properties, not mechanisms. In: Raskin V, ed. Proc. of the 2004 Workshop on New Security Paradigms. New York: ACM, 2004. 67−77.

[5] T. Rauter, A. Holler, N. Kajtazovic, and C. Kreiner, “Privilege-based ¨ remote attestation: Towards integrity assurance for lightweight clients,” in Proceedings of the 1st ACM Workshop on IoT Privacy, Trust, and Security. ACM, 2015, pp. 3–9.

[6] Luo, W., Liu, W., Luo, Y., Ruan, A., Shen, Q., & Wu, Z. (2016). Partial Attestation : Towards Cost-Effective and Privacy-Preserving Remote Attestations.2016 IEEE Trustcom/BigDataSE/ISPA.IEEE,2016,pp.152 – 159

[7] Abir Awad; Sara Kadry; Brian Lee; Gururaj Maddodi; Eoin O'Meara.Integrity Assurance in the Cloud by Combined PBA and Provenance.2016 10th International Conference on Next Generation Mobile Applications, Security and Technologies (NGMAST).2016,pp,127-132

[8] Merkle RC. Protocols for public key cryptosystems. In: Proc. of the IEEE Symp. on Security and Privacy. Washington: IEEE Computer Society, 1980. 122−134.

[9] Merkle RC. A certified digital signature. In: Brassard G, ed. Proc. of the 9th Annual Int’l Cryptology Conf. on Advances in Cryptology. Heidelberg: Springer-Verlag, 1989. 218−238.

[10] Blum M, Evans W, Gemmell P, Kannan S, Naor M. Checking the correctness of memories. In: Proc. of the 32nd Annual Symp. on Foundations of Computer Science. Washington: IEEE Computer Society, 1991. 90−99.

[11] Gassend B, Suh GE, Clarke D, van Dijk M, Devadas S. Caches and hash trees for efficient memory integrity verification. In: Proc. of the 9th Int’l Symp. on High-Performance Computer Architecture. Washington: IEEE Computer Society, 2003. 295−306.

[12] Xu, Z.-Y., He, Y.-P., & Deng, L.-L. (2011). Efficient Remote Attestation Mechanism with Privacy Protection. *Journal of Software, 22*(2), 339–352.

2017 International Conference on Computer Network, Electronic and Automation (ICCNEA 2017)

Key Point Detection in Images Based on Triangle Distribution of Directed Complex Network

Qingyu Zou
College of Electrical and Information Engineering
Beihua University
Jilin, China
zouqingyu2002@126.com

Jing Bai *
College of Electrical and Information Engineering
Beihua University
Jilin, China
jlbyj@163.com

Jianwen Guan
College of Electrical and Information Engineering
Beihua University
Jilin, China
408088149@126.com

Weiliang Sun
College of Electrical and Information Engineering
Beihua University
Jilin, China
2460696609@qq.com

***Abstract*—Key point detection is still a challenging issue in pattern recognition. With the recent developments on complex network theory, pattern recognition techniques based on graphs have improved considerably. Key point detection can be approached by community identification in directed complex network because image is related with network model. This paper presents a complex network approach for key point detection in video monitoring image, which is both accurate *and fast. We evaluate our method for square and subway* station video monitoring images. Results show that our algorithm can outperform other traditional method both in accuracy and processing times.**

Keywords-Key point detection; Complex network; Community identification

I. INTRODUCTION

Key points are a set of pixels in an image which are characterized by a mathematically well-founded definition, which are rich in terms of information content. They are commonly used as local features in many image applications such as content-based image retrieval or object recognition[1]. The best-known key point detectors include Moravec algorithm, Harris algorithm, genetic-programming algorithms, and so on. The Moravec algorithm defines the corner strength of a point as the smallest sum of squared differences between the point patch and its adjacency patches. The Harris detector computes the locally averaged moment matrix using the image gradients, and then combines the eigenvalues of the moment matrix to compute the strength of each corner. The genetic programming methods automatically synthesize image operators aimed to find the key points in an image using fitness functions which measure the stability of the operators through the repeatability rate, and also promote the uniform dispersion of detected points [1, 2].

With the development of complex networks theory, key point detection based on graph theory is a new research hotspot in the field of image recognition in recent years. This method maps the image to the weighted and directed graph, treats the pixel as the node, and obtains the best key point of the image with the optimal shear criterion. This method essentially transforms the key point detection problem into the optimization problem. It is a kind of point-to-clustering method, and it has a good application prospect for data clustering. At present, the research of key point detection based on graph theory mainly focuses on the following aspects: (1) the design of optimal shear criterion; (2) the spectral method is used for detection; (3) the design of fast algorithm [3, 4, 5, 6, 7].

Communities are defined as subsets of highly inter-connected nodes, relatively and sparsely connected to nodes in other communities[8]. They have intrinsic interest because they may correspond to functional units within a complicated system. In this paper, we introduce a novel approach to computing the key points of an image using complex network community identification. We construct a weighted and directed complex network model to represent each image that gives some valuable information about the location of the key points. The nodes in the network model are the super pixels of the image, which allows a drastic reduction in the number of representative pixels in an image, and the edge is the relationship between the super pixels, according to variations of the intensity, color and other parameters. Then the key points could been identified through the communities identification of the image network model.

II. SPARSE NETWORK MODEL OF IMAGE

The description of the image as a complex network model is based on the complex network theory for image recognition of the premise and foundation. The complex network model is divided into unqualified network, undirected weight network, directed network and directed weight network.

We use the directed and weighted network to describe the image. The super pixels in the image are taken as the nodes of the network, and the luminance difference $|I(p_i)-I(p_j)|$ of the two pixels $[i, j]$ is used as the weight between the two points $w_{i,j}$, when the image is grayscale. When the image is a color map, the edge weight $w_{i,j}$ between node *i* and *j* is the RGB European distance of them.

$$W_{i,j}=\frac{\sqrt{\left(R_i\text{-}R_j\right)^2+\left(G_i\text{-}G_j\right)^2+\left(B_i\text{-}B_j\right)^2}}{\sqrt{3}} \tag{1}$$

Where R, G, and B are the RGB values of the pixels, respectively. The direction of the edge in the network is the brightness of the pixels point to the brightness of small pixels.

In order to make the uniform network into a complex network, we calculate the Euclidean distance between any two pixels. Then set the radius threshold *r* and delete the links between the nodes in the network where the distance are greater than r. The final weight of the complex network links w_{ij} is given by the formula (2) [9, 10, 11, 12]. The complex network model of the image is constructed as shown in Fig. 1. The characteristics of the complex network model of image are shown in the Table 1.

$$w_{ij}=\begin{cases}\left|I(p_i)-I(p_j)\right|, & \left(dist(p_i,p_j)\le r\right)\\ 0, \quad \left(dist(p_i,p_j)>r\right)\end{cases} \tag{2}$$

TABLE I. THE CHARACTERISTICS OF THE COMPLEX NETWORK MODEL

Characteristic	**Value**
Node	3775
Link	610561
Average in-out-degree	4132.2
Average in-closeness	0.0463
Average out-closeness	0.0928
Average betweenness	4635.7
r	35

Degree measure the number of edges connected to it in a network. Although degree is a simple centrality measure, it can be very illuminating. Closeness measures the mean distance from a node to other vertices. This quantity takes low values for vertices that are separated from others by only a short geodesic distance on average. Such vertices might have better access to information at other vertices or more direct influence on other vertices. Betweenness measures the extent to which a node lies on paths between other vertices. Vertices with high betweenness centrality may have considerable influence within a network by virtue of their control over information passing between others[14].

III. NETWORK COMMUNITY IDENTIFICATION

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. *Triangle Vertexes Weightiness*

In general, the simple building blocks of complex networks are a small structure of several nodes called motif[15]. Network motifs are small subgraphs that can be found in a network statistically significantly more often than in randomized networks. Among the possible motifs, the simplest one is the triangle which represents the basic unit of transitivity and redundancy in a graph, see Fig. 2.

As shown in Fig. 2, there are 13 triangle cases at most, including 39 vertexes, in an arbitrary directed network. We compare all three vertexes one another for each triangle T_i and merge the code of vertexes had the same place. Then, there are 30 special vertexes for triangles, encoded from 1 to 30 in Fig. 2. We assign different weights w_i to different vertexes *i*, because some complex triangles contain the simple triangles, such as triangle 11 contain 1. We assign higher weights to the vertexes whose are not affected by other vertexes, and lower weights to depend on other vertexes. The w_i is calculated using a function as follows:

$$w_i=\frac{TC_i}{\max(TC_i)} \tag{3}$$

where TC_i means the number of vertexes affected by vertex *i*. We consider that each vertex affects itself. For instance, for vertexes 1, TC_1=2, since it affects vertexes 25 and itself; similarly, TC_6 = 3, since vertex 6 affected vertexes 17, 20 and itself.

B. *Triangle Degree*

The number of triangles that the node touches is the triangle degree of it. The eigenvalue of nodes is used to measure the structure characteristic of nodes in network. The eigenvalue of node *i* is defined as follow：

$$\xi\left(n_i\right)=\sum_{t\in\phi}n_t^i w_t \tag{4}$$

where n_i is a node in the network n_t^i is the number of triangle vertexes connected node *i* in the position *t*.

C. *Process of Network Community Identification*

- Construct the complex network model of image.
- Calculate the triangle degree of each node in a complex network model.
- Hierarchical clustering according to node degree.
- Calculate the best value of each segmentation.

- Extract key points from the image according to the division of the communities in the complex network model.

IV. Results and Discussion

In order to verify the effectiveness of the proposed algorithm, we use this method for square and subway station video monitoring image. In two experiments we compare our method with Harris algorithm, which is a well know key point detection approaches.

A. Experiment 1-Square

This first experiment, we employed a square image. The characteristics of the complex network model of square image are shown in the Table 2. Fig. 3 shows the key point detection results of square image produced by our algorithm (a) and Harris algorithm (b). The clustering results of square image by our algorithm obtained a total of 6 communities. Our algorithm based on community identification producing 55 key points, which 41 nodes are valid nodes. Harris algorithm producing 138 key points, which 24 nodes are valid nodes.

TABLE II. The characteristics of the square image complex network model

Characteristic	Value
Node	885
Link	3344
Average in-out-degree	128.21
Average in-closeness	0.0566
Average out-closeness	0.0658
Average betweenness	148.22
r	35
Maximum Q_{od}-value	0.1736

B. Experiment 2-Subway Station

In this experiment, we employed a subway station image. The characteristics of the complex network model of square image are shown in the Table 3. Fig. 4 shows the key point detection results of subway station image produced by our algorithm (a) and Harris algorithm (b). The clustering results of subway station image by our algorithm obtained a total of 7 communities. Our algorithm based on community identification producing 51 key points, which 30 nodes are valid nodes. Harris algorithm producing 6 key points, which 3 nodes are valid nodes.

TABLE III. The characteristics of the subway station image complex network model

Characteristic	Value
Node	732
Link	24146
Average in-out-degree	1280.9
Average in-closeness	0.0489
Average out-closeness	0.0513
Average betweenness	640.66
r	50
Maximum Q_{od}-value	0.1040

V. Conclusion

Is this paper we presented a feasible algorithm based on complex networks and pixels for the key point detection in video monitoring images. This algorithm construct the complex network model of image based on the characteristics of image super pixels, firstly. Then identify the key nodes according to triangle distribution features of image directed complex network model by community identification. We showed that it provides accurate key point of video monitoring images within very low processing times.

Acknowledgment

This work is supported mostly by the Jilin Province department of education Science and Technology Plan Projects. (Grant NO. 201658), partially by the Beihua University Dr. Scientific Research Fund (Grant NO. 2015557) and partially by the Jilin Province in the Science and Technology Development Plan Project (Grant NO. 20170520057JH and NO. 20160418038FG).

References

[1] R. Criado, M. Romance, et al., “Interest point detection in images using complex network analysis,” Journal of Computational and Applied Mathematics, vol. 236, Dec. 2012, pp. 2975-2980.

[2] Z. X. Wu, X. Lu, et al., “Image edge detection based on local dimension: A complex networks approach,” Physica a-Statistical Mechanics and Its Applications, vol. 440, 2015, pp. 9-18.

[3] K. Jieqi, L. Shan, et al. “A complex network based feature extraction for image retrieval,” IEEE Press, 2014.

[4] J. Tang, Y. Chen, et al., “Image Modeling and Feature Extraction Method Based on Complex Network,” Computer Engineering, vol. 39, 2013, pp. 243-247.

[5] Y. Chen, J. Tang, et al., “Image representation and recognition based on directed complex network model,” Advances in Intelligent Systems and Computing, vol. 212, 2013, pp. 985-993.

[6] Q. Li, J. Ye, et al., Interest point detection using imbalance oriented selection, Pattern Recognition, vol. 41, 2008, pp. 672-688.

[7] A. R. Backes, D. Casanova, et al., “Texture analysis and classification: A complex network-based approach,” Information Sciences, vol. 219, 2013, pp. 168-180.

[8] M. E. J. Newman, “Communities, modules and large-scale structure in networks,” Nature Physics, vol. 8, 2012, pp. 25-31.

[9] J. d. A. S. Wesley Nunes Gonçalves, Odemir Martinez Bruno, “A Rotation Invariant Face Recognition Method Based on Complex

Network.” In Proceedings of the 15th Iberoamerican Congress on Pattern Recognition, 2010: 426-433.

[10] A. R. Backes and O. M. Bruno, “Shape classification using complex network and Multi-scale Fractal Dimension,” Pattern Recognition Letters, vol. 31, 2010, pp. 44-51.

[11] R. Criado, M. Romance, et al., Interest point detection in images using complex network analysis, Journal of Computational and Applied Mathematics, vol. 236, 2012, pp. 2975-2980.

[12] R. Criado, M. Romance, et al., A post-processing method for interest point location in images by using weighted line-graph complex networks, International Journal of Bifurcation and Chaos in Applied Sciences and Engineering, vol. 22, 2012, pp. 1250163.

[13] W. d. Nooy, A. Mrvar, et al., “Exploratory social network analysis with Pajek.” New York: Cambridge University Press, 2011.

[14] M.E.J.Newmen, “Networks: An Introduction.” New York: Oxford University Press, 2010.

[15] O. Shoval and U. Alon, SnapShot: Network Motifs, Cell, vol. 143, 2010, pp.326-U158.

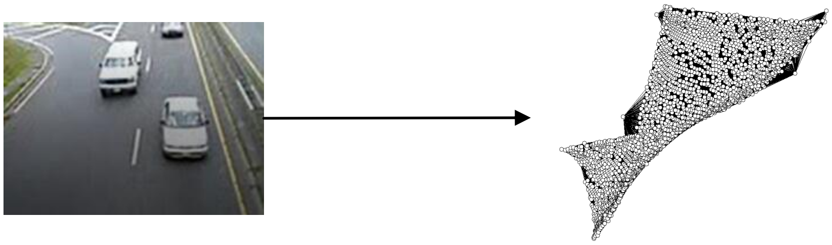


Figure 1. Complex network model of image. The picture has been done using the software Pajek[13]

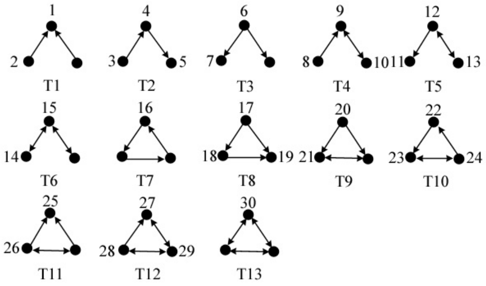
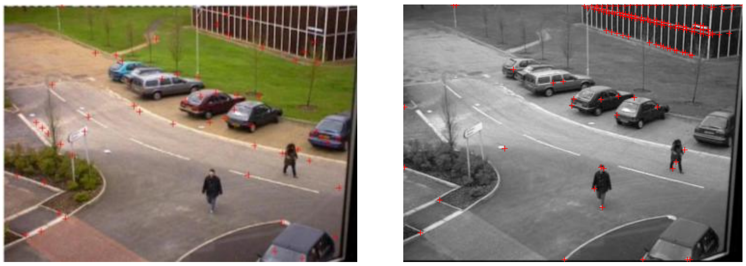


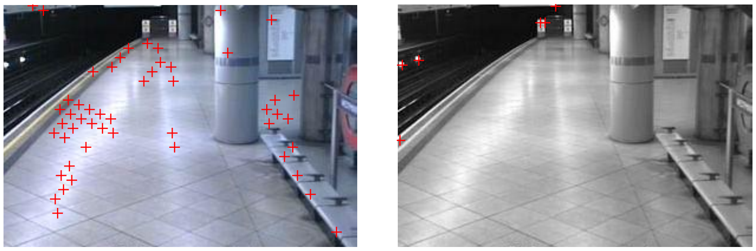
Figure 2. List of all 13 types of triangles



(a) Our algorithm

(b) Harris algorithm

Figure 3. Key point detection in square image



(a) Our algorithm

(b) Harris algorithm

Figure 4. Key point detection in subway station image

Research on Combination Forecasting Model of Mine Gas Emission

Liang Rong
College of Computer Science and Technology
Xi'an University of Science and Technology
Xi'an, China
E-mail: liangr79@163.com

Jia Pengtao
College of Computer Science and Technology
Xi'an University of Science and Technology
Xi'an, China
E-mail: pengtao.jia@gmail.com

Chang Xintan
College of Safety Science and Engineering
Xi'an University of Science and Technology
Xi'an, China
E-mail: changxt@xust.edu.cn

Dong Dingwen
College of Safety Science and Engineering
Xi'an University of Science and Technology
Xi'an, China
E-mail: 289425071@qq.com

Abstract—This paper focuses on the effective analysis of the mine gas emission monitoring data, so as to realize the accurate and reliable mine gas emission prediction. Firstly, a weighted multiple computing models based on parametric *t*–norm is constructed. And a new mine gas emission combination forecasting method is proposed. The BP neural network model and the support vector machine were used as the single prediction models. Finally, genetic algorithm and least square method were used to determine the parameters of t-norm in the combination forecasting model, and realized the optimal combination of single models. The experimental analysis shows that the new model has less error than BP neural network model and support vector machine model in the evaluation indexes. It can be concluded that the new combined forecasting model is more suitable for the coal mine gas emission forecast.

Keywords-Mine gas emission; Combination forecasting; Parametric t–norm; LS-SVM; BP neural network

I. INTRODUCTION

Mine gas emission is a very complex dynamic phenomenon; it is comprehensive affected by the Coal seam burial depth, coal seam thickness, mining intensity, the original gas content and other factors. The interaction of various factors make the change of gas emission is dynamic non-linear characteristics. Gas emission forecast is the important reference for mine ventilation system design and gas control management. According to the collected data, establishing a suitable forecasting model to forecast gas hazards, during the early phase of gas accumulation and concentration overrun. Then it can guide the relevant department to take preventive measures to prevent the occurrence of disaster accidents, to ensure the safety of coal production.

There are four main ways to forecast the current gas emission:

1) Statistical analysis. According to received lots of gas emission actually from the previous mine production and mining depth of the data, using the laws of mathematical statistics promotes the forecast to the new mine.

2) Coal seam gas content method. According to the gas content of coal seam and the residual gas content in coal after mining to calculate relative gas emission.

3) Source calculation method. According to the law of gas emission, calculated separately coal mining face, excavation face, mining area and mine gas emission.

4) Analogy method. According to the known mine gas emission to predict the mine with the same or similar geological conditions and mining conditions [1].

With the rapid development of science and technology, especially the development of mathematics and computer technology, the original prediction methods and application fields have been expanded, and some new or further optimized prediction methods have emerged. It mainly uses regression analysis[2,3], BP neural network[4-6], support vector machines[7,8], wavelet analysis[9,10]etc. These models have made some predictive effects when predicting the amount of gas emission. However, in view of the application of a single prediction model, the modeling mechanism is different and will have some limitations in different degrees. Therefore, the generalization ability is poor.

Since the t-norm can improve the generalization ability of the system, based on the prediction model of individual gas emission, the author proposes a combined forecasting model based on parametric *t*-norm to forecast the amount of gas emitted by a number of factors.

II. WEIGHTED MULTIVARIATE COMPUTING MODEL BASED ON T-NORM

A. Parametric T-norm

Definition 1[11] Set binary operator $t(x,y)$ is the binary operation of [0,1]×[0,1]→[0,1]. The $t(x,y)$ that satisfies the following conditions is t-norm:

1) Boundary conditions: $t(0,y)=0, t(1,y)=y$;

2) Monotonic: $t(x,y)$ *on* x,y *monotonically increasing;*

3) Associative law: $t(t(x,y),z)=t(x,t(y,z))$ *;*

4) Commutative law: $t(x,y)=t(y,x)$.

Gas emission is the important parameters of prevention and control of gas explosion and gas outburst warning. In view of the complexity of the application area, the concept of *t*-norm needs to be appropriately expanded. Since the computation and the operation itself can be continuously changed, a parametric *t*-norm is proposed. The parameterized *t*-norm is the extend of the And operation and the Or operation, and satisfies the four conditions in the *t*-norm definition. The form of the *t*-norm with parameters is determined by the parameters, using different *t*-norm generation functions, various t-norms and parametric *t*-norm can be generated. So, *t*-norm has good generalization ability.

B. Multivariate Parameterization T-norm

From the definition of *t*-norm, it is a binary operator. In application, the *t*-norm must be extended to a multivariate function.

Theorem 1 the *t*-norm can be generalized to a multivariate function.

By the Associative Property $t(x_1,x_2,...,x_n)=t(t_{n-1}(x_1,x_2,...,x_{n-1}),x_n)$, Theorem 1 is established.

C. Weighted Multivariate Computing Model

The current *t*-norm arithmetic model describes an equal condition in an ideal state. And many of the influential factors in the actual complex system generally have different weights. So we introduce the weight parameter for the parameterized *t*-norm as follows:

Theorem 2 Set *f(x)* is the generation function of *t*-norm, then

$$G(x_1,x_2,...,x_n,\alpha_1,\alpha_2,...,\alpha_n) = f^{-1}(\max(f(0),\sum_{i=1}^{n}\alpha_i f(x_i)-1)) \qquad (1)$$

is a weighted multivariate computing model based on parametric *t*-norm. Among that, $x_i \in R^+(i=1,2,...,n)$, α_i is the weight of $f(x_i)$, $\alpha_i \in [0,1]$ and $\sum_{i=1}^{n}\alpha_i = 1$.

When the function *f (x)* changes, the model generates a new operator cluster, and more variability, so that the model has good generalization ability. Set *f*(*x*)=*x*^{p}, so the generated weighted multivariate computing model from (1) is:

$$G(x_1,x_2,...,x_n,\alpha_1,\alpha_2,...,\alpha_n) = f^{-1}(\sum_{i=1}^{n}\alpha_i x_i^p - 1) \qquad (2)$$

Among them, $p \in (-\infty,0)\cup(0,+\infty)$, the composite forecasting model is constructed according to (2).

III. Modeling Method of Gas Emission Prediction

A. BP Neural Network Modeling Method

There is a nonlinear relationship between gas emission and many influencing factors. Therefore, BP neural network can be used to predict the gas emission [12]. In this paper, Single prediction model of BP neural network for gas emission quantity is based on three layers BP neural network, the input layer is composed of 6 nodes, which is used to input different kinds of influence factors; and the output layer has 1 node, which is for the prediction value of the gas emission quantity. During the comparison of multiple training models for gas emission prediction models, and finally determined to adopt one hidden layer, the incentive function uses logarithm sigmoid function. Hidden node number using for 4 has better training effect. The initial weight set among the (-1,1). Convergence error is set to 1e-3.Construction of BP neural network predictive model structure is showed by Fig .1.

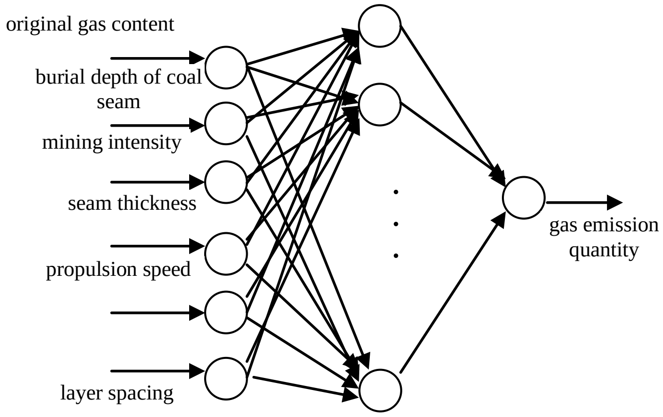


Figure 1. The neural network prediction model for mine gas emission

B. SVM Regression Model Modeling Method

Support Vector Machines (SVM)is a new approach of learning machine based on VC dimension theory and structural risk minimization principle of statistical learning theory. It shows unique advantages in the solution of limited samples, non-linear and high-dimensional pattern recognition, learning and local minimum and other issues. The basic thought of using SVM to estimate the regression function is through non - linear mapping φ, and input the data x for the space mapped to a high-dimensional feature space, then in this high-dimensional space for linear regression.

Least Squares Support Vector Machines (LS-SVM) is one of the method bases on SVM, and replaces the traditional SVM with the least-square method system, using quadratic programming method to solve the problem of pattern recognition. It changes the quadratic programming problem of the algorithm of the original SVM to system of linear equations by constructing a new quadratic loss function. So that it can effectively reduce the computational complexity. So, the paper uses LS-SVM as the single prediction model to forecast the gas emission.

The step of establishment of LS-SVM prediction model of gas emission is as following:

1) Given a training set made by a sample data of N gas emission factors $\{x_k, y_k\}_{K=1}^{N}$ *,inputs data* $x_k \in R^m$ *,outputs data* $y_k \in R$ *,the function fitting problem can be described as the following optimization problem:*

$$\min_{w,e} J(w,e) = \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{k=1}^{N} e_k^2 \qquad (3)$$

$$S.\ T.\ y_k = w^T \phi(x_k) + b + e_k (k = 1,2,\ldots,N) \qquad (4)$$

In (4), $\phi(x_k)$ is used to map input data from space R^m to high dimensional feature space R^{mh} ,and $w \in R^{mh}$ is the weighted variable. $\gamma > 0$ is penalty factor to use for adjust error; $e_k \in R$ is error variable, $b \in R$ is offset value.

2) Using the way of Lagrange to solve this optimization problem, that is:

$$\begin{gathered} L(\omega,b,\zeta,\alpha,\gamma) = \\ \frac{1}{2} \omega_g \omega + c \sum_{k=1}^{n} \zeta_k^2 - \sum_{k=1}^{n} \alpha_k (\varphi(x_k) + b + \zeta_k - y_k) \end{gathered} \qquad (5)$$

Among that , $\alpha_k (k = 1,2,\ldots,N)$ is Lagrange multiplier. From $\frac{\partial L}{\partial \omega} = \frac{\partial L}{\partial \zeta} = \frac{\partial L}{\partial \alpha} = \frac{\partial L}{\partial \gamma} = 0$,it can calculate equation：

$$\omega = \sum_{k=1}^{n} \alpha_k \phi(x_k), \sum_{k=1}^{n} \alpha_k = 0, \alpha_k = c^{\zeta} k \qquad (6)$$

Kernel function $K(x_i, x_j) = \phi(x_i) g^{\phi}(x_j)$ is symmetric functions satisfying Mercer condition. According to (6) and the constraint (4), the optimization problem can be transformed into solving linear equations.

$$\begin{bmatrix} 0 & 1 & \wedge & 1 \\ 1 & K(x_1,x_1)+1/c & \wedge & K(x_1,x_1) \\ M & M & M & M \\ 1 & K(x_1,x_1) & \wedge & K(x_1,x_1)+1/c \end{bmatrix} \begin{bmatrix} b \\ \alpha_1 \\ M \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} 0 \\ y_1 \\ M \\ y_1 \end{bmatrix} \qquad (7)$$

3) The LS-SVM fitting model is obtained

$$y(x) = \sum_{k=1}^{N} \alpha_k K(x_k, x) + b \qquad (8)$$

Thereinto, α_k is the support vector, α_k and b can be calculated according to the training sample data.

The training of LS-SVM model is mainly to solve the linear equations(7). When using LS-SVM model to forecast, just need to calculate the kernel function between the training samples and the tested samples $K(x_i, x_j)$, not involving the concrete form of the function $\phi(x_k)$.

4) Selected kernel function. Choosing kernel function is an important part of building model. Because Radial Basis Function(RBF) has good learning ability and wide convergence domain, RBF is chosen as the kernel function of the model. That is:

$$K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / \sigma^2) \qquad (9)$$

Among that , σ is kernel function.

C. Combination Forecasting Modeling Method

The combined forecasting modeling method is a combination of two or more predictive methods to predict the same prediction problem. Theoretical research and practical application, the combination forecast makes full use of the advantages of each individual forecasting model, having the ability to adapt to the development of future forecast conditions. It can enhance the stability of prediction and improve the accuracy of prediction[13, 14].The key of the combined forecasting model lies in the generalization ability of the model. The combined forecasting model can be described as follows:

Assuming that the actual observed value of a predicted object at Time t is $y(t)(t = 1,2,...,m)$,there are n feasible predictive methods for this prediction problem, the corresponding prediction models are f_1 , f_2 ,…, f_n ,and the predicted values are $\hat{y}_i(t)(t = 1,2,...,m; i = 1,2,...,n)$, that is $\hat{y}_i(t) = f_i(t)$ .Weighted combination forecasting problem can be described as:

$$\hat{y}(t) = F(\hat{y}_1(t), \hat{y}_2(t), ..., \hat{y}_n(t), \alpha_1, \alpha_2, ..., \alpha_n) \qquad (10)$$

Among that, combination forecast value is $\hat{y}(t)(t = 1,2,...,m)$, F is a way of combination. The purpose of using (10) is to make combination forecast value $\hat{y}(t)$ have better effect than single prediction $\hat{y}_i(t)$.

The combined forecasting model is based on BP neural network theory and Least Squares Support Vector Machines theory, the predicted values are: $\hat{y}_{BP}(t)$ and $\hat{y}_{LS-SVM}(t)$ .Combined with the characteristics and advantages of each single prediction model, the different weights α and $1-\alpha$ are assigned to the individual forecasting models for the combined model. Assuming the F in (10) is (2), so the mathematical expression of Combination Forecasting Model based on Parametric T-norm(CFMPT) is:

$$\hat{y}(t) = ((\alpha \hat{y}_{BP}(t)^p + (1-\alpha) \hat{y}_{LS-SVM}(t)^p - 1)^{1/p} \qquad (11)$$

The genetic algorithm and the least squares are combined to estimate the parameters.

Because of the objectivity and inevitability of prediction error, there are some error between predicted value $\hat{y}(t)$ and actual value $y(t)$. Set error:

$$E = ((\hat{y}(t) - y(t))^2)^{1/2} \tag{12}$$

To minimize the error E as the objective function of the genetic algorithm, the two parameters of CFMPT model are obtained.

Genetic algorithm is an adaptive global optimization search algorithm to simulate the genetic and evolutionary processes of biological organisms in the natural environment. In view of its global search ability, CFMPT parameter estimation module use genetic algorithm. Set the following genetic algorithm parameters for parameter optimization: the initial population is 20; Using binary encoding, the number of encoding is 8; Select the operation using a stochastic uniform distribution model; the crossover operation uses a distributed cross. The mutation operation uses Gaussian function variation[15].

IV. Experiment and Result Analysis

A. Experimental Design

The main factors influencing the gas emission (m^3/ min) include: burial depth of coal seam(m), mining intensity(average daily output, t/d), seam thickness(m), propulsion speed (average daily progress, m/d), original coal seam gas content(m^3/t) and layer spacing(m).These are used as input. Gas emission is used as output. Using the No. 1-15 data as the training sample, which from the Table. V in the [1] named monitoring data of absolute gas emission in a coal mine, the No. 16-18 is used as prediction sample to forecast.

The experimental steps are as follows:

1) *For comparison, the sample is normalized and the data is normalized to [0,1].The normalized formula is as follows:*

$$norm(x_i) = \frac{x_i - \min(X)}{\max(X) - \min(X)} \tag{13}$$

2) *Training the individual prediction methods BP and LS-SVM on the training set prediction results* $y_{BP}(t)$ *and* $y_{LS-SVM}(t)$ *are obtained on the test set.*

3) *Using CFMPT model: genetic algorithm is used to estimate the parameters in the training set, prediction results of CFMPT is reached in the test set.*

4) *All regression results are subjected to an anti-normalization process as follows:*

$$y_i = \hat{y}_i \times (\max(X) - \min(X)) + \min(X) \tag{14}$$

5) *Evaluation by evaluation index.*

B. Result and Analysis

Using CFMPT model to forecast, the results are shown as Table. I, according to the genetic algorithm to find the optimal Equation (11). That is, the parameter of the combination forecasting model is $\alpha = 0.982$, $p = 3.649$. For comparison, the results of the single prediction model and the results of [1] and [9] are also listed into the table. The results of the comparison on the error evaluation index are shown in Table. II.

TABLE I. Comparison of Test Results

Sample No	Real value	Improvement BP[1]	Wavelet Analysis [9]	BP (Article)	LS-SVM (Article)	CFMPT
1	4.06	3.99	4.02	4.0616	4.6759	4.0708
2	4.92	4.58	4.98	4.9328	4.7882	4.9301
3	8.04	8.08	8.09	8.0828	7.4782	8.0387

TABLE II. Comparison of the Errors of Various Algorithms (%)

Sample No	Improvement BP[1]	Wavelet Analysis[9]	BP (Article)	LS-SVM (Article)	CFMPT
1	1.72	0.99	0.04	15.17	0.27
2	6.91	1.21	0.26	2.68	0.20
3	0.50	0.62	0.53	6.99	0.02
Aver-age Errors	3.043	0.94	0.277	8.28	0.163

It can be seen from the Table. II, when it uses various prediction models to forecast the gas emission, The average relative error of improved BP[1] is 3.043%.The average relative error of the BP prediction model proposed in this paper is 0.277%.The average relative error of the wavelet neural network[9] is 0.94%.The average relative error of LS-SVM model in this article is 8.28%.The average relative error of CFMPT model is the lowest, as 0.163%.

Prediction results show, BP neural network, wavelet neural network and LS-SVM all can be used to forecast gas emission. But for small sample, number of hidden layers of BP neural network will lead to a large difference between the predicted results. For example, improved BP neural network in [1] and using the BP neural network in this article, all lead to differential production , because of the difference between the training method and the number difference of hidden layers. But the number of hidden layers of the current selection has no theoretical guidance. most of them are selected by experience, and the number of training is too much, it is easy to be fitting situation. Regardless of BP model, wavelet neural network model or LS-SVM model, because of the limitation of single model, it is hard to be sued into all cases. In another words, the generalization ability is weak. The combined forecasting model compensates the shortcomings of the single model, learns from each other, and obtains the best forecasting effect.

The single prediction model in the CFMPT prediction model can adopt any linear, nonlinear model. Due to the variability of CFMPT, it is single event prediction model can be established. And the CFMPT model is most suitable for predicting the characteristics of time series data.

V. Conclusion

A combined forecasting model of gas emission based on parametric t-norm is proposed. CFMPT model has the ability to change, generalization ability. It can use the way of training to find the best combination model for the characteristics of data sets. That overcomes the limitation of the weak ability of small sample of individual forecasting method.

Experimental results show that the result from CFMPT model is better than the result from single prediction model, which verifies the effectiveness of the method. Moreover, it has strong reliability for time series prediction, and has good practical application value.

Acknowledgment

Shaanxi Provincial Department of Education Scientific Research Project (No.15JK1454).

Xi'an Science Program Project (No.CX1519(3)).

References

[1] LI Hongbiao, Research on the Prediction of Gas Emission by Using Neural Network. Kunming: Kunming University of Science and Technology,2008.

[2] YE Zhen-ni, HOU En-ke, DUAN Zhong-hui, HE Dan, PAN Yi, “Prediction for gas emission quantity of the working face in Guojiahe coal mine,” Journal of Xi’an University of Science and Technology, vol. 37, Jan. 2017, pp. 57-62, doi:10.13800/ j.cnki.xakjdxxb. 2017.0110.

[3] Wang Jiangrong, “Gas Emission Prediction Model Based on Genetic Algorithm and Fuzzy Multivariate Linear Regression Analysis,” Industry and Mine Automation, vol. 39, Dec. 2013, pp. 34-38, doi:10.7526/j.issn.1671-251X.2013.12.009.

[4] WANG Sheng-quan, LIU Bo-gen,ZHANG Zhao-zhao, FAN Qi, FENG Hai, “Prediction of Gas Emission Quantity of Mining Faces Based on Genetic BP Neutral Network Optimal Model,” Journal of Xi’an University of Science and Technology, vol. 32, Jan. 2012, pp. 51-56, doi:10.13800/j.cnki.xakjdxxb.2012.01.022.

[5] Gao Baobin, Pan Jiayu, “Based on PLS associated with BP neural network for different-source gas emission prediction model of working face,” Journal of Hunan University of Science & Technology(Natural Science Edition) , vol. 30, Dec. 2015, pp. 14-20, doi: 10.13582 / j.cnki.1672-9102.2015.04.003.

[6] SONG Da-yong, SONG Guo-liang, CHENG Gao-fei, YANG Xiao-min, “Research on the prediction of gas emission from tunneling roadway by using neural network,” Shaanxi Coal, 2012(4), pp. 141-142,128.

[7] FU Hua, XIE Sen, XU Yao-song, CHEN Zi-chun, “Study on MPSO-WLS-SVM-based Mine Gas Emission Prediction Model,” China Safety Science Journal, vol. 23, May. 2013, pp. 56-57, dio: 10.16265/j.cnki.issn1003-3033.2013.05.020.

[8] FU Hua, YU Xiang, LU Wanjie, “Prediction of Gas Emission Based on Hybrid Algorithm of Ant Colony Particle Swarm Optimization and LS-SVM,” Chinese Journal of Sensors and Actuators, vol. 29, Mar. 2016, pp. 373-377, doi:10.3969/j.issn.1004-1699.2016.03.012.

[9] CaoHongmin, ZhangYulin, JiangYongpeng, DingChengwei. “Mine Gas Gushing Quantity for Ecasting Based on Wavelet Neural Network,” Computer Applications and Software, vol. 26, Jul. 2009, pp. 168-170.

[10] ZHANG Wensheng, “Research on Gas Emission Pattern Recognition at Coal Heading Face Based on Wavelet Analysis,” Beijing: China University of Mining & Technology,,2016.

[11] Masaharu MIZUMOTO, “Pictorial Representation of Fuzzy Connectives, Part 1: Cases of T-norms, T-conorms and Averaging Operators,” Fuzzy Sets and Systems, 1989,31:217-242.

[12] WeiYin-shang,LiuYun-fei, “To Study on Forecasting of Mine Gas Emission Based on The Monte Carlo method improved BP Neural Net,” Coal Engineering, vol. 46, No.12, 2014, pp. 84-86,89, doi:10. 11799/ce201412028.

[13] TIAN Shui-cheng, WANG Xi, WANG Li, CEN Xiao-qian, YUAN Xiao-fang, “Combination forecast method of gas emission amount based on improved AHP,” Journal of Xi’an University of Science and Technology, vol. 32, Jan. 2012, pp. 25-29, doi:10.13800/j.cnki.xakjdxxb.2012.01.018.

[14] AO Pei, MOU Long-hua, “Load combination forecasting based on power grid with environmental characteristics,” Journal of China Coal Society, vol. 36, Sep. 2011, pp. 1575-1580, doi:10.13225/j.cnki.jccs.2011.09.025.

[15] JIA Peng-tao, DENG Jun, “Coal Mine Gas Concentration Combination Prediction Model Based on Universal Average Operation,” China Safety Science Journal, vol. 22, Jun. 2012, pp. 41-46.

2017 International Conference on Computer Network, Electronic and Automation (ICCNEA 2017)

WLAN Based Wireless Self-organization Link: Research and Realization

Yunhai Guo
Engineering Research Center of Digital Audio &Video
Ministry of Education
Communication University of China
Beijing, China
E-mail: guoyunhai@cuc.edu.cn

Zhengxiang Li
Engineering Research Center of Digital Audio &Video
Ministry of Education
Communication University of China
Beijing, China
E-mail: lizhengxiang@cuc.edu.cn

Abstract—**Communication is one of essential elements of our life. In some cases, we need wireless links to support data transmission such as television relay. But the existing GPRS and other network often need the support of power transmission line. And in the remote areas, the transmission rate of traditional network is very slow. In this paper, the development and realization of wireless self-organization link based on the wireless router and WLAN, managed by OpenWrt open source system is presented. It implements a strategy for Wireless Self-organization Link system, which can effectively solve the problem of automatic relay between wireless routers, prevent the establish of loopback in two routers, make multiple routers connect automatically, and set up a high-speed wireless data link in AP+Client mode.**

Keywords-WLAN; Wireless Self-organization Link; OpenWrt; Ad Hoc; WSOL

I. INTRODUCTION

SINCE 1971, Norman Abramson, a professor at the University of Hawaii, developed the world's first wireless computer communication network, ALOHA net. The Wireless communication gradually spread and the improvement of the data quantity is a higher requirement for the signal transmission. However, the laying of the signal cable is often a very complicated process, and in some cases, such as mountainous area or the disaster area, the cable is not allowed to lay. At this time, the signal transmission requires the use of wireless links. The existing Internet, GPRS network and so on often need the support of power transmission line, and in the remote areas, the transmission rate of traditional network is very slow, not enough to support the high-definition video signal transmission.

At present, the Wireless Mesh Network (WMN) which based on the IEEE802.11s and Wireless Sensor Network (WSN) are gradually becoming mature [1], [2]. Last year, the RSSI-based swarm robots passing narrow corridor based on Mesh network, and the adaptive video protection in large scale peer-to-peer video streaming which based on WMN has been developed [3], [4]. As the main problem of solving the mesh network interconnection, the data rate requirements are often slow [5]. Moreover, the design of the special wireless link transmission based on the router has not appeared.

In this paper, based on the OpenWrt system, a Wireless Self-organization Link (WSOL) strategy is implemented. This strategy is based on IEEE802.11 protocol which can be used to establish the wireless connection between the nodes effectively. And this method can implement the self-assembling and self-constructing of the link node, and effectively prevent the generation of self-loop.

The design can effectively improve the coverage of the wireless link, and provide support for the high speed data transmission.

II. WSOL SYSTEM OVERVIEW

The structure of WSOL used in this paper is based on WLAN and IEEE802.11 protocol [6], which is “one-by-one” link structure. As shown in Figure 1.

Source Terminal and Relays are routers use OpenWrt system and user program. This architecture is divided into three parts: Source Terminal (S), Relay (C), and User Terminal (U). The Source Terminal is directly connected to the Central Treatment System and is only used as an Access Point (AP); Relay uses AP+Client mode on the one hand as the client access to AP, on the other hand to build its own AP for the next Relay or User Terminal access. The User Terminal can be any device with the function of wireless access, such as mobile phones, cameras, etc.

In this structure, under the control of the program algorithm, the user deploys several Relays and a Source Terminal between User Terminal and Central Treatment System. S and Central Treatment System connected by cable. A wireless link between U and Central Treatment System can be automatically generated when we turn on the power supply of each router. And when a relay node failure, just replace the fault router, open a new router can re-establish links.

We use the OpenWrt system in the router. OpenWrt is a highly extensible GNU/Linux distribution for embedded devices (typically wireless routers). Its core is a Linux system with only the most basic functions. OpenWrt provides a fully writable file system and software package management, which is ideal for writing programs and even modifying systems according to the project requirements [7]. Therefore, we use OpenWrt as the system of routers and make use of its mature 802.11 protocol stack, write a series of program control routers to achieve the function of WSOL.

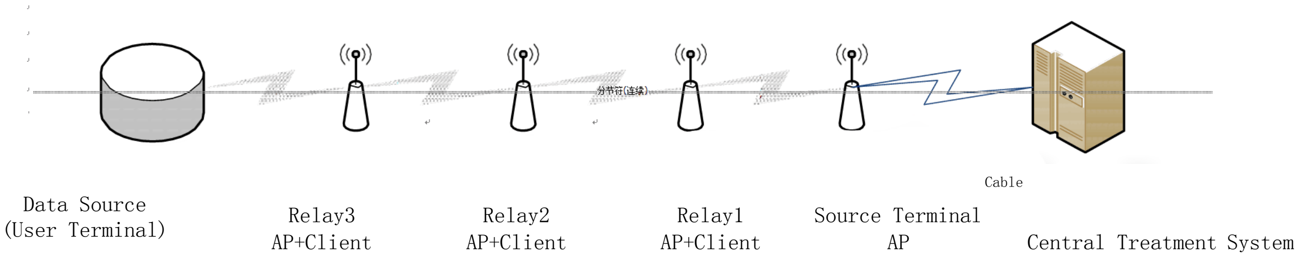


Figure 1. WSOL structure hart

UCI (Unified Configuration Interface) is a small utility written in C and is intended to centralize the entire device's configuration running OpenWrt. Examples are the main network interface configuration, firewall settings, wireless settings and remote access configuration. OpenWrt central configurations are split into several files located in the
/etc/config/ directory [8]. We can edit the configuration files with a text editor or modify them with the command line by program “uci”. UCI configuration files are also modifiable through various programming APIs like Shell or C.

OpenWrt systems are widely used in various router brands, so this scheme can work on many routers. And this system can improve the speed and stability by updating the hardware or the built-in protocol.

III. SYSTEM DESIGN

A. *Summary of nodes*

The WSOL system includes the following key nodes: Source Terminal (only one), User Terminal (only one), and Relay (N nodes). The workflow of each node is as follows:

1) Source Terminal: The Source Terminal is directly connected to the Central Treatment System. After boot, the terminal launch its AP firstly, and then turnoff DHCP.

2) User Terminal: The User Terminal can be any device with the function of wireless access, such as mobile phones, cameras, etc. After boot, connect to the strongest AP around terminal.

3) Relay: Because the Relay uses AP+Client mode on the one hand as the client access to AP, on the other hand to build its own AP for the next Relay or User Terminal access, the “Discovery, Decision and Connect” algorithm is mainly applied to Relay. There are 3 kinds of connection of Relay (base on Fig.1) :

a) Relay1 connected to Source Terminal

b) Relay3 connected to Relay1

c) Insert Relay2 between Relay3 and Relay1 when Relay3 is connected to the Relay1 or Relay3 is disconnected from Relay1 because of the distance between Relay3 and Relay1 is too far.

B. *The “Discovery, Decision and Connect” algorithm*

- Step1: Relay1 startup firstly, then, scan APs around the Relay1, connect to the strongest AP. (Relay1 connected to Source Terminal).
- Step2: Relay3 startup firstly, then, scan APs around the Relay3, connect to the strongest AP. (Relay3 connected to Relay1).
- Step3: (At this time, insert Relay2 between Relay3 and Relay1) After Relay3 connected to the Relay1, when the signal strength of the newly inserted Relay2 is greater than that of the original connection (the signal strength of Relay1 detected by Relay3) 20dBm, the connection is going to be changed. When Relay3 making changes, Relay3 disconnect original connection firstly, then, wait 10s (prevent Relay2 gear into Relay3) and connect to Relay2 (the strongest AP).
- Step4: After Relay2 connected to an AP, check if it can communicate with Source Terminal. If not, break off the connection and select another AP to join. As shown in Fig. 2.

It should be noted that, a certain distance between Relays is necessary. The signal strength difference needs to reach 20dBm. The routers use static IP address.

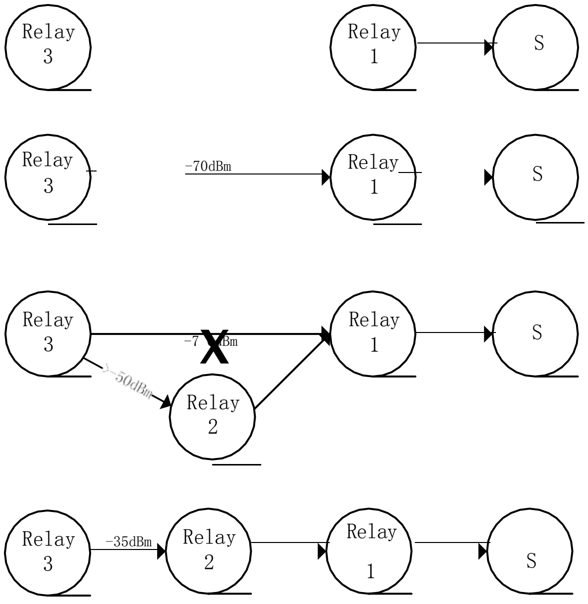


Figure 2. Algorithm flow of Relay

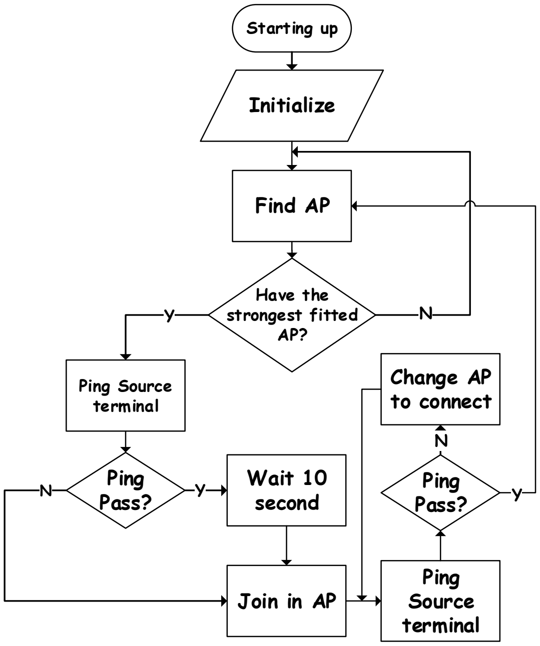


Figure 3. System flow chart of Relay

IV. IMPLEMENTATION

In this section, the implementation details will be explained in five parts: Setup AP and the initial work, find routers, choose routers, join in AP, decide and change router. The flow chart of relay terminal's behavior is as shown in Fig. 3.

A. Setup AP and the initial work

For the Relay, the first work after boot is turned on its AP and get ready to connect another device. Specifically, we must turn on the AP function of the device, and configure the channel, power, interface, SSID, encryption, LAN IP, then establish new interface "WWAN", configure firewall, shut down DHCP of LAN interface, and restart network service in the end.

Because the router uses the Linux kernel, there are many network configuration tools in Linux we can take advantage. The network interface, wireless functions and firewall functions are easily to configure by "uci" command. If these command lines are written into a Shell file, they can be run automatically.

B. Find Router

After the AP and the initial configurations are completed, the Relay needs to know other routers around it and find the fitted AP to connect. In this paper, we use a uniform method to name the SSID of AP. For example, we use a string "openwrt" connect up to the last segment of IP address. Such as [1] "openwrt12" means the LAN IP of this AP is "192.168.12.12". The advantage of this method is we can select the appropriate AP in complexity wireless circumstance.

C. Choose Router

After getting information of surrounding APs, this information needs to be selected again by router. The best AP will be selected which signal strength is the strongest one. Then, the information of the best AP will be written into file "bestrouter.txt" and this file will be supplied to AP access program.

D. Join in AP

After determining the best AP, the connection job is about to begin. The information of this AP (SSID, MAC address) will be read into a program, which can use "uci" command to configure the IP address, gateway address, interface mode and MAC address of interface "WWAN". Then, restart network, the connection job is done.

In this paper, the AP+Client mode is used in Relay. We utilize the interface LAN as the interface which can be accessed by another router, and the interface "WWAN" as the interface to connect another router. The LAN mode is "ap", the WWAN mode is "sta".

E. Change Router

Aiming at the third connection cases of Relay, we need to detect the current communication between the Relay and Source Terminal. To be specific, after Relay connecting to the best AP, check the communication with Source Terminal, break off the connection and select another AP to join if can't communicate with Source.

After all these steps, a reliable link will be established between Source Terminal and the last Relay. We tested this link on a board based on *AR9341*[2], and *iperf*[3] program on OpenWrt system to send data between Relays.

After link relay, no significant fluctuations in average bandwidth, multi-hop link runs well.

V. CONCLUSION

In this work, a WSOL system using OpenWrt build-in system has been presented and tested. By compiling C and Shell programs and running in wireless routers, it is able to establish data links between two points automatically. The average bandwidth between two points is no significant fluctuations. However, this design also has its shortcomings. Because the link update is based on the communication with the Source Terminal, the update speed is slow, encounter a collision of the test signal will bring wait time. So this work is not applicable for mobile wireless network. In the future work, we plan to improve the algorithm to enhance the mobility and optimize the link detection scheme to reduce the re-build connection time.

ACKNOWLEDGMENT

This work was supported by national sci-tech support plan, digital media technology development. (2015BAK05B00, 2015BAK05B01).

REFERENCES

[1] IEEE Standard for Information Technology--Telecommunications and information exchange between systems--Local and metropolitan area networks--Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 10: Mesh Networking," in IEEE Std 802.11s-2011 (Amendment to IEEE Std 802.11-2007 as amended by IEEE 802.11k-2008, IEEE 802.11r-2008, IEEE 802.11y-2008, IEEE 802.11w-2009, IEEE 802.11n-2009, IEEE 802.11p-2010, IEEE 802.11z-2010, IEEE 802.11v-2011, and IEEE 802.11u-2011) , vol., no., pp.1-372, Sept. 10 2011.

[2] Yick, J., Mukherjee, B., & Ghosal, D. (2008). Wireless sensor network survey. Computer Networks, 52(12), 2292–2330.

[3] Hattori, K., N. Tatebe, T. Kagawa, Y. Owada, L. Shan, K. Temma, K. Hamaguchi and K. Takadama. "Deployment of Wireless Mesh Network Using Rssi-Based Swarm Robots Passing Narrow Corridor by Movement Function Along Walls." Artificial Life And Robotics 21, no. 4 (2016): 434-442.

[4] Ghaeini, Hamid Reza, Behzad Akbari, Behrang Barekatain and Alicia Trivino-Cabrera. "Adaptive Video Protection in Large Scale Peer-to-Peer Video Streaming over Mobile Wireless Mesh Networks." International Journal Of Communication Systems 29, no. 18 (2016): 2580-2603.

[5] Chakraborty, Sandip and Sukumar Nandi. "Data Rate, Path Length and Network Contention Trade-Off in Ieee 802.11s Mesh Networks: A Dynamic Data Rate Selection Approach." Computer Networks 91, (2015): 225-243.

[6] Serrano, P., P. Salvador, V. Mancuso and Y. Grunenberger. "Experimenting with Commodity 802.11 Hardware: Overview and Future Directions." Ieee Communications Surveys And Tutorials 17, no. 2 (2015): 671-699.

[7] Anonymous, “OpenWrt”, https://openwrt.org/, accessed by 27 April 2017.

[8] Anonymous, “UCI”, https://wiki.openwrt.org/doc/uci, accessed by 27 April 2017.

Egwra: Qos Routing Algorithm In Wireless Mesh Networks Based On Evolutionary Game Theory

Yan Weiguang
Department of Information Technology
Hunan Women's University
Changsha, China
E-mail: 173873656@qq.com

Pan Xianmin*
Department of Information Technology
Hunan Women's University
Changsha, China
E-mail: 15974239623@139.com
*The corresponding author

Abstract—**This paper applies the theory of Evolutionary Game to QoS routing algorithm for wireless mesh networks which can not only improve the performance of traditional QoS routing protocols but also be able to reduce the cost of the routing algorithm.**

Keywords-Wireles Mesh Networks; Routing protocol; Qos Routing; Evolutionar y Game Theory; Routing algorithm

I. INTRODUCTION

As a new type of wireless network, wireless mesh network connect mesh nodes through wireless links to construct a dynamic topology, self-organizing and multi-hop wireless interconnected network. Compared with the traditional single-hop wireless networks, it can extend coverage, enhance robustness, reduce deployment cost and increase capacity. Compared to the traditional switched networks, wireless Mesh network cabling between nodes removed needs, but still has a distributed network provides redundancy and re-routing. In the wireless Mesh network, if you want to add a new device, simply plug in the power on it, it can automatically configure itself, and determine the best multi-hop transmission path. Add or mobile device, the network topology changes can automatically find and automatically adjust the traffic routing in order to obtain the most efficient transmission path. Wireless mesh network's business is usually gathered in the Mesh Router or Gateway, easily lead to local network congestion, making it difficult to maintain network globally optimal routing. Thus routing protocols must be able to adapt to this situation so as to provide better QoS for the users. So, the research of wireless mesh network routing protocol is of great significance. For example, Zhou, Hao[1] focuses on QoS routing with bandwidth constraint in multi-radio multi-channel WMN, and proposes a new multimetric and a QoS routing protocol MMQR. The routing metric has two advantages. First, it replaces the transmission rate of ETT with available bandwidth so that the nodes with light load are more likely to be selected. Second, it takes the channel diversity into account and assigns a weight to each link according to the channels of links within the range of three hops. Sun, Xuemei[2, 3] proposes a QoS routing algorithm based on culture-particle swarm optimization algorithms. The algorithm uses the dual-evolution mechanism of culture algorithms and achieves further improvement on global optimum location mutation particle swarm optimization algorithms (MPSO) by introducing the concept of inertia weight. Traditional QoS routing algorithms like the achievements listed above only considered single objective performance parameters, such as delay bound or bandwidth limitations, or static multi-objective constrained situation. WMN can not meet the needs of General Requirements for some business like dynamic multi-objective performance, such as delay, bandwidth and Frequency congestion.

This paper applies the theory of Evolutionary Game to QoS routing algorithm for wireless mesh networks which can not only improve the performance of traditional QoS routing protocols but also be able to reduce the cost of the routing algorithm.

II. DESCRIPTION OF EGWRA

A. Network Model

In this paper, we consider the WMN with three-tier architecture. The WMN consists of wireless routers and mobile stations. The wireless router serves as an access point in IEEE 802.11 for a number of mobile stations within its transmission range. Moreover, the wireless router also acts as the relay node to transmit packets between it and its neighboring wireless routers. All the wireless routers are located at fixed sites, and they are interconnected via wireless links to form a mesh backbone. The communication of wireless links usually follows the IEEE 802.11 standard [3]. Some wireless routers are designated as gateways to connect the mesh backbone with Internet. The gateways also hold the physical wires to connect with Internet. For the mobile station, it is the end device (e.g. PDA, laptop, etc.) in the WMN, which can randomly move within the WMN. Its packet transmission is through one or more wireless routers in the mesh backbone, and then via a gateway to the Internet. For the packet from Internet to a mobile station, this packet transmission is via the above reversed route.

B. Channels of IEEE 802.11

The IEEE 802.11b/g based WMN model is extensively used in most of WMN literature. This claim is also made in [5]. The IEEE 802.11b/g operates in the 2.4 GHz spectral band, which is split into 11 channels for use in the US. Only three channels (1, 6, and 11) are non-overlapping. The remaining channels are partially overlapping. For each non-overlapping channel, its frequency band is not overlapped with the other two. In contrast, the frequency band of each partially overlapping channel is overlapped with some others. In this paper, we exploit all available (non-overlapping and partially overlapping channels) in IEEE 802.11b/g to perform channel assignment.

C. Related Work

In WMNs, most of existing channel assignment algorithms only use non-overlapping channels to perform channel assignment. In [4-5], the experiment results demonstrated that the throughput of the WMN can be enhanced when all available (partially overlapping and non-overlapping) channels are appropriately used in channel assignment. In [4-5], the authors have also quantified the interference effects of IEEE 802.11b/g in terms of the channel separation and the corresponding interference factor. The channel separation is the difference between two used channels for two nodes. The interference factor is defined as the ratio (I/R) of the interference range (I) to the transmission range (R).

In addition to measuring the channel interference, the research work of [5] also utilizes all available channels of 802,11b/g to propose a channel assignment algorithm called the multichannel multicast (MCM) algorithm. In this algorithm, it gives a channel selection function IR ( ). If channel c is to be assigned for node p, all the neighbors (all the nodes within the transmission range) of node p are first found. Then, the channel separation between node p and each neighbor is a key parameter to respectively calculate the corresponding function IR ( ). Next, all the values of the function IR ( ) are added. After calculating the total value of IR ( ) for each available channel, the channel with the minimum total value of IR ( ) will be assigned for node p. From the above description, we can know that the main idea of the MCM algorithm is to reduce the interference effect between a node and its neighboring nodes. The channel assignment sequence follows the level order. The node on the higher level is earlier than the node on the lower level to be assigned a channel with less interference.

In [5], the channel assignment algorithm is called the minimum interference multi-radio multicast (M4) algorithm. In the M4 algorithm, two-hop interference range is considered in the channel selection to avoid the hidden terminal problem [7]. In the MCM algorithm [4], the interference range is only considered within one hop. For the channel selection metric of the M4 algorithm, it aims to balance the channel separations for the nodes within two hops. This can avoid some node pairs incurring sever interference due to using small channel separations. However, the interference cost is not minimized in the given channel selection metric. In addition, the channel assignment sequence is not discussed in the M4 algorithm.

To compare the above two channel assignment algorithms with the algorithms just using non-overlapping channels, the MCM and M4 algorithms have great improvement in the throughput of the WMN.

D. Summary of Evolutionary Game Theory

Non-cooperative game theory is the decision-making in a distributed environment, the analysis of individual utility maximization Player for the optimal policy choice. Evolutionary game non-cooperative game, a branch of a game strategy for further analysis of game populations in a long-term stability. Evolution of the Nash equilibrium (all Player of the optimal strategy) with groups of stability, that is executed when the other Player balanced strategy, any Player can not be balanced by a unilateral departure from the strategy for more effective; Meanwhile, the implementation of a balanced strategy can reveal the individual proportion of total population.

As the novel achievement in the research field of non-cooperative game theory[8-10], the research on evolutionary game theory attracts great attentions in not only academy but also industry field. Integration of evolutionary game theory, economics and evolutionary biology of rational thought, no longer human model into the game super-rational side, that the human is usually achieved through trial and error method of game balance, and biological evolution is common, the choice The balance is the balancing process to achieve a balanced function, and thus the historical, institutional factors and the balancing process are some of the details of the game will affect the choice of multiple equilibria[11].

Set the evolution game located in an N-node MANET, any node with M, that except the node i other than the collection. denotes the data packets generated by source node i which is called i's group. Data between source and destination nodes transmit a complete data service is called a session; the node mobility will lead to changes in network topology, the completion of a session requires multiple routing paths to create different groups.

The other parameter M trust set up in all the main components of the set of strategies for the real number field on an interval X, strategy by the probability density distribution f (x) to characterize and design the fitness function is a continuous function in domain , which is the environmental parameters, as Environment, the probability density. Note the probability density function of the composition of all the set. The evolution of the network configuration software, game, evolutionary selection is linked with fitness. When given the definition of the conditions of environmental parameters selection operator and the average selection operator , the environment, the average fitness function.

According to the statements listed above. The evolution game for WNN can be defined[12]. So that G = (I, P, U). Where: I = {Ni, N-i} for the Player collection, Ni said node i, and Ni indicates that the network nodes in addition to all the nodes outside i; P is the strategy set P = {pi, p-i}, 1 packet transmission , refused to transmit a message to 0. U is the utility function, the utility function with Ui (si, s-i) to represent. It proceeds B (si, s-i)> 0 from the i (the correct transmission) and costs C (si, s-i) <0 (energy consumption) of two parts. Node i can be the watchdog or other monitoring and feedback mechanisms, that message transmission to the next node if there are Byzantine errors. Whether the benefits or costs i have a strategy with all relevant participants, this is a strategy game or action nodes are the embodiment of interaction, about the Byzantine fault-tolerant WMN classic prisoner's dilemma problems and to link the two adjacent Between nodes Ni and Nj prisoner's dilemma shown in Table 1.

TABLE I. THE GAME MATRIX BETWEEN LABOR NODES IN WNN

	successful	**failure**
The packet reachs node N_j	(*B* ,*C*)	(0 , *C*)
The packet not reachs node N_j	(*B*, 0)	(0, 0)

E. Working Steps of EGWRA

According to the definition of evolution game , we can design the working steps of EGWRA as follows.

```
/* Step 1：Finding Available EGWRA-Neighbor */
  TCBTC-N(u) □ Φ；TD(u) □ Φ；p(u) □ p0；α = π/3
  while ( p(u) < P and gap-α(TD(u)) )
  begin
  bcast (u, p(u), (Hello, p(u)) )
  u receives Ack (ack, p(u)) message from v
  if v's game score is more than the threshold
  u calculates the direction of discovered neighbor v dir(u,v), the transmitting power and the direction determines the neighbor v( p(u), dir(u, v) )
  TCBTC-N(u) = TCBTC-N(u) ∪ { v | discovered neighbor v }
  TCBTC-D(u) = TCBTC-D(u) ∪ {dir(u, v) | discovered neighbor v }
  Pow(u) = Pow(u) ∪ { p(u, v) | discovered neighbor v }
  p(u) □ Increase(p(u))
  end
/* Step 2：Finding Available DT-Neighbor */
  k is the upper bound of node degree, k = 6
  Sort all qualified neighbors found in Step1 in order of increasing distance or power
  Pow = { p1, p2, p3, ……, pm }，p1≤ p2≤ p3≤ ……≤pm
  while (payoff value>threshold)
  begin
    for( i=1; i ≤m; i++ )
    begin
    u transmits with power pi, 1≤i≤m
    draw a perpendicular bisector between u and the node corresponding to the power pi
      end
    end
    TDT-N(u) = TDT-N(u) ∪ { v | v ∈ TCBTC-N(u) and v has corresponding Voronoi Edge }
    TDT-D(u) = TDT-N(u) ∪ { v | v ∈ TCBTC-N(u) and v has corresponding Voronoi Edge }
/* Step 3：Filling the α -gap */
  Sort TCBTC-N(u) and TDT-N(u) in order of increasing direction
  if gap-α ( TDT-N(u) )
    then α □ the smaller direction of the gap
    β □ the larger direction of the gap
    if dp, dq, dr ∈TCBTC-N(u) and dp ≤ dq ≤ dr and dp =α , dr =β
    then dq is dropped in Step 2 and can fill the α-gap
    TN(u) = TDT-N(u) ∪ { v | the direction of v is dq }
    TD(u) = TDT-D(u) ∪ { dir(u,v) | the direction of v is dq }
    Pow(u) = Pow(u) ∪ { p(u, v) | the direction of v is dq }
/* Step 4：Edge Removal */
    Suppose node v, w∈N(u)
    send(u, p(u, v), relation(v, w), v)
    recv(u, relation(v, w), v)
    if relation(v, w) is "Y" and | TN(u) | - 1 ≥ 3
    then TN(u) =TN(u) - {v}
    Procedure gap-α(TD(u))
    Suppose TD(u) = {d1,d2,d3,…,dn}
    Sort directions in TD(u) in increasing order
    TD(u) = {dk1,dk2,dk3,⋯,dkn}, dk1≤dk2≤dk3≤⋯≤dkn, 1≤ki≤n, 1≤i≤n
    for( i=1; i ≤n; i++ )
      begin
    if dki+1 - dki ≤ 2π /3 then
    continue
    end
```

III. SIMULATION EXPERIMENT OF EGWRA

A. The Design of Simulation Experiment

We use the simulation software made by a Ns 2.30 pairs relay cooperation strategy to construct the experiment. The topology of experiment WNN can be described like figure 1, which divided into two parts, mesh network and underlay network.

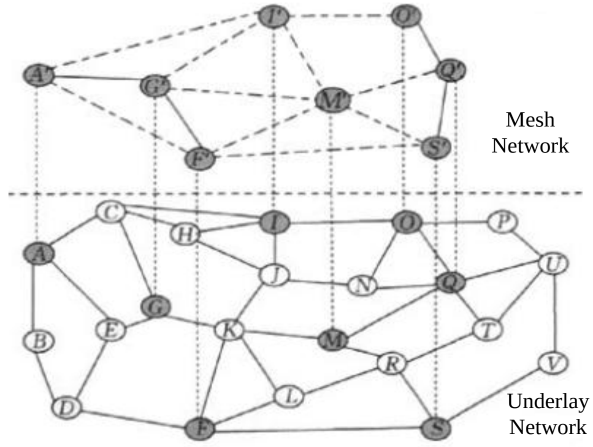


Figure 1. Simulation result of EGWRA.

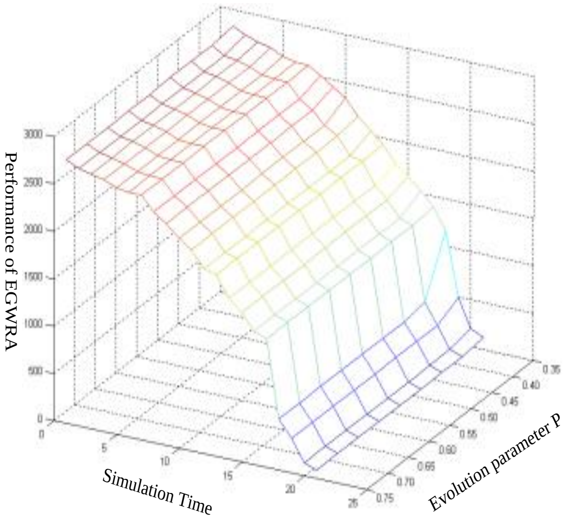


Figure 2. The topology of experiment network.

Simulation using routing EGWRA. Collaboration defined level index node for the network in any of NARMeaning the actual relay nodes and the number of packets following the packet in which the number of requiredVolume ratio. NAR ∈ [o, 1], the greater the value of a node's NAR. Help the higher the degree of collaboration involved in relay. Simulation scenario is set as follows. number of nodes N for a 50, each section Point speed in the 10 ~ 20m/ s were evenly distributed within the mobile rangeWai as a 500m × 500m, each node in a session immediately after the end of Initiate another session carved, and the duration of each session 5-10swere uniformly distributed within. These settings will ensure that each node in the network Interaction between the higher the probability of the relay. Each session is a fixed-speed transmission Rate of 1 Mb / s data stream, packet size is 512byte. Each simulation True continuous 900 s, 50 times in the implementation of the statistical average demand.

B. *Result Analysis of Simulation*

According to the parameter we start the experiment and the simulation result can be described like figure 2. At fixed intervals of time, t = 1-75, movement occurs by updating the speed and direction of each MN. We have chosen the tuning parameter used to vary the randomness. The speed and direction are chosen from a random Gaussian distribution with mean equal to zero and standard deviation equal to one. For a random chosen instant t in total simulation time, our proposed approach extracts 21 MNs as a stable core in terms of mobility which is marked with a read dot.

This extraction-core based on mobility promotes reliable links between MNs. Therefore, the QoS-aware routing find a path that require QoS through this core, the reliability and lifetime of QoS are more guaranteed in time.

From the experiment we can reach a conclusion that the performance of routing is greatly improved and the cost is not increased very much.

IV. CONCLUSION AND FUTURE WORKS

With the rapid development of wireless technology, some related technologies like mesh network play more and more important roles in working and living process. This paper applies the theory of evolution game to the Qos Routing algorithm and proposes a novel method called EGWRA which integrated multi-step game process into routing decision making process and the performance of routing is greatly improved and the cost is not increased very much.

REFERENCES

[1] .Zhou, H., et al. A new multi-metric QoS routing protocol in wireless mesh network[C]. 2009. Wuhan, Hubei, China: Inst. of Elec. and Elec. Eng. Computer Society.

[2] .Sun, X., C. Li, and M. Zhang. A QoS routing algorithm based on culture-particle swarm optimization in wireless mesh networks[C]. Chengdu, China: IEEE Computer Society.

[3] Traore Soungalo, Li Renfa, Zeng Fanzi, "Evaluating and Improving Wireless Local Area Networks Performance", International Journal of Advancements in Computing Technology, vol. 3, no. 2, pp. 156-164, 2011.

[4] G. Zeng, B. Wang, Y. Ding, L. Xiao, and M.W. Mutka, "Efficient Multicast Algorithms for Multichannel Wireless Mesh Networks", IEEE Trans. Parallel and Distributed Systems, vol. 21 no. 1, pp. 86-99, Jan. 2010.

[5] H. L. Nguyen and U. T. Nguyen, “Chanel Assignment for Multicast in Multi-channel Multi-radio Wireless Mesh Networks,” Wiley InterScience. Wireless Communications and Mobile Computing, vol. 1 no. 4, pp. 557-571, April 2008.

[6] .Romdhani, L. and C. Bonnet. Cross-layer QoS routing framework for wireless mesh networks[C]. 2008. Athens, Greece: Inst. of Elec. and Elec. Eng. Computer Society.

[7] J. So and N. Vaidya, “Multi-Channel MAC for Ad Hoc Networks: Handling Multi-Channel Hidden Terminals Using A Single Transceiver”, Proc. ACM MobiHoc, pp. 222-233, 2004. 、4.

[8] Zhang, Y., Z. Fang, and Z. Bu. Evolution game analysis of flowage strategy of science and technology persons of Nanjing[C]. 2007. Nanjing, China: Inst. of Elec. and Elec. Eng. Computer Society.

[9] Chen, Z., S. Yang, and Y. Cao. Profit allocation in mobile commerce value chain based on the evolution game theory[C]. Chongqing, China: IEEE Computer Society.G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[10] Madhusudan Singh, Sang-Gon Lee, Whye Kit Tan, Jun Huy Lam, "Impact of Wireless Mesh Networks in Real-time Test-bed Setup", Advances in Information Sciences and Service Sciences, vol. 3, no. 9, pp. 25-33, 2011.

[11] V. Mhatre, “Enhanced Wireless Mesh Networking for ns-2 Simulator”, In ACM SIGCOMM Computer Communication Review, pp. 69 – 72, 2007.

[12] Nori Mohammed AL-Kharasani, Zuriati Ahmad Zukarnain, "Performance Evaluation of Routing with Load-Balancing in Multi-Radio Wireless Mesh Networks", International Journal of Advancements in Computing Technology, vol. 5, no. 2, pp. 64 -71, 2011.

.

Design of a WSN Node for Rice Field based on Hybrid Antenna

Huaqiang Chen
College of Electronic Engineering
South China Agricultural University
Guangzhou 510642, P.R.China
243591735@qq.com

Baoxia Sun
College of Electronic Engineering
South China Agricultural University
510642, P.R.China
sunbaoxia@126.com

Weixing Wang(Correspondence author)
College of Electronic Engineering
South China Agricultural University
Guangdong Engineering Research Center for Monitoring Agricultural Information
weixing@scau.edu.cn

Jiangpeng Weng, Fenglian Tie
College of Electro+nic Engineering
South China Agricultural University
Guangzhou 510642, P.R.China
vmishiwjpv@163.com, 65517386qq.com

Abstract—Aim at the problems existing in the information monitoring of the farmland environment such as the limited energy, low system stability and large monitoring area, a WSN node for rice field based on hybrid antenna is designed to realize the real-time on-line monitoring for the environmental parameters of rice fields in the network. As for the hardware, the node uses a STM32F103VET6 as a processing core, and a WLK01L39 RF chip is used in wireless communication module, while the sensor module is composed of the air temperature and humidity sensor, light intensity sensor and soil moisture sensor. As for the software, uC/OS-II is applied as an operational system to realize multitask scheduling running. The sensor node applies a mechanism as sleeping and waking up work modes to reduce power consumption. The current consumption of sensor nodes is 0.024mA under the sleeping mode, 32.32mA under the data collection, 26.25mA under data transmission and 21.95mA under the operating mode. The results of a long time networking experiment indicate that the average PLR (Packet Loss Rate) of network is 0.76%. In conclusion, the design of sensor node system is suitable for the real-time and stable monitoring of rice field.

Keywords-Hybrid antenna, rice field, low power consumption, node design

I. INTRODUCTION

In our country, the planting area and production of rice occupy the first position among the main grain crops [1,2,3]. However, some problems have restricted the development of rice industry, such as the low technology content, the backward production technology and small production scale. Wireless sensor network (WSN), which is considered to be one of the most important technologies in the 21st century[4], is a kind of new information acquisition and processing technology that can be applied in the field of agricultural. WSN is a fully-distributed system that typically composed of a large number of sensor nodes with specific functions. The advantages of simple deployment, intensive layout, low cost and no need for on-site maintenance provide convenience for the data acquisition of environmental science researches. Combining with the farmland environment parameter sensor [5], the real-time on-line monitoring of large area farmland environment parameter can be realized, which has great significance to the accelerating of the agricultural informatization in our country and the prediction on the degree of the damage of rice drought and diseases of insect pests. There have been many examples of successful application at home and abroad [6-114].

In accordance with the broad monitoring environment, adequate sunlight and few architectural barriers of rice fields, regarding the air temperature and humidity, the light intensity and the water temperature of the soil as monitoring objectives, this paper designs a node for rice filed based on the hybrid antenna. In order to meet the application requirements of stable real-time on-line monitoring in rice field environment, the wireless sensor network node is designed to realize the energy self-supply.

II. HARDWARE DESIGN OF WIRELESS SENSOR NODES

A. Directional Antenna

In the application of wireless sensor nodes, omnidirectional antenna which has 360 ° horizontal radiation patterns is suitable for the multipoint communication. In the case of a need for sending and receiving information in the whole network, the omnidirectional antenna can guarantee effective reception among nodes and the horizontal movements of wireless node will not influence communication, which make it easy to install and manage. In the designed environmental information real-time monitoring system for rice fields based on WSN, the sensor nodes randomly dispersed in the network monitoring area. Due to the low requirements for its communication distance and direction, the omnidirectional antenna is equipped. The main task of cluster head nodes is to collect data of sensor nodes and realize the reliable communication over a long distance with gathering node, thus, the directional antenna is equipped and its direction is set towards the gateway nodes. Due to the limited length, this paper illustrates the hardware design method of the system with the representative example of the collecting node design.

B. Hardware Structure of Sensor Node

Nodes constitute a complete wireless sensor network in the form of self-organization. The environmental monitoring information of rice field was sent through single hop to the cluster head nodes, from the cluster head nodes to the gateway nodes, then to the monitoring center by the network. In this way, it can achieve the monitoring environment parameters acquisition of rice field [15]. The acquisition node consists of a processor module, a power supply, a solar charging module, a sensor module and a wireless radio frequency module. The WSN node structure is shown in Fig. 1.

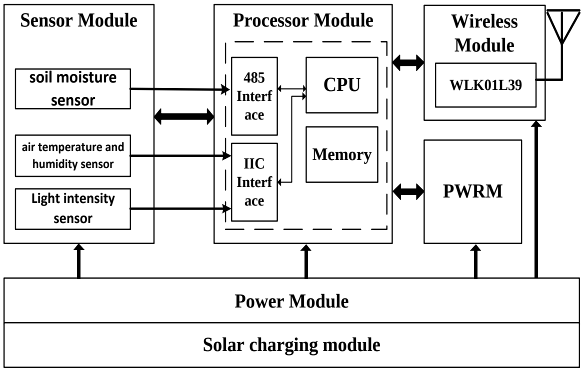


Figure 1. The structural of WSN node

C. Processor Module

In order to meet the design requirements of superior performance and low power consumption, the microprocessor uses STM32F103VET6 produced by ST Microelectronics. It has high command execution efficiency and strong anti-jamming capability. What's more, it supports standby sleep and has the functions of equipment management, task scheduling, data fusion processing, etc.

D. Sensor Module

The sensor module is mainly responsible for the collection of air temperature and humidity, light intensity, and soil temperature and humidity in the rice field. According to the accuracy of the data parameters and the needs of field application , the design uses the HTU21D temperature and humidity sensors, the light intensity sensor BH1750FVI, as well as the soil moisture temperature sensor SMTS-II-485 produced by French Humirel Company.

III. SOFTWARE DESIGN OF WSN NODES

The performance of the wireless sensor network nodes and the stability of the network are directly affected by the software design. Considering the characteristics of the long cycle, strong disturbances and small quantity of data in single transmission, a software system based on the uC/OS-II is designed according to the hardware platform mentioned above . Due to the use of embedded operating system, the design greatly improves the availability of the software to prolong the lifetime of nodes.

A. Design of Communication Protocol among Nodes

Limited by the node energy, network computing and processing capacity and other resources, the design of WSN communication protocol is particularly important. The design reduces the cost of energy and data transmission delay by means of routing algorithm, time synchronization algorithm and standby wake-up mechanism, and thus the allocation of WSN resources can be optimized. The routing protocol designed in this paper is improved partly based on the LEACH protocol. This system is aimed at the large rice field so that the design uses the clustering routing algorithm. Cluster head is randomly selected in each round. Data packets of nodes in a cluster are sent to the cluster head and data packets between clusters can be forwarded to the gateway node through multi hop of cluster heads. Cluster head nodes first check their own record after receiving a new packet. If the packet has been forwarded, they will not be forwarded again. In this way, the waste of energy caused by forwarding the same data packets can be avoided. The structure of packets sent to the cluster head by sensor nodes and packets sent to the gateway node by cluster head is shown in Figure. 2.

Data Flag	Cluster Num	Node Num	Sensor 1	Sensor 2	Sensor 3	Battery Voltage	Jump Num	CRC

Figure 2. Sensor node packet structure

B. Time Synchronization Algorithm Design

After the gateway nodes obtain the current time and the standby time from the remote server, it will process the information packet of the synchronous time, and then send to the cluster heads and nodes in clusters through broadcast. Each node in clusters receives the synchronous packet, which will be parsed into the current time and the alarm clock time, then set the RTC clock register before entering the low power standby mode. The structure of synchronization time information packet is shown in Figure3.

Syn Flag	Current Time	Alarm Time	CRC

Figure 3. Synchronization time packet structure

C. Standby Wake-Up Mechanism

In the network, each node has its local clock. After the system starts, the network will provide a common time stamp for local clocks of all nodes and it will be on standby or be awakened at the right time. By means of this standby mechanism, the whole network can save the cost and reduce the energy consumption.

D. Application Design

The version of small embedded real-time system in acquisition nodes is V2.91 uC/OS-II. The system is responsible for preemptive real-time multi tasks scheduling. Tasks communicate with each other through a signal. The whole system is divided into several parts, including initialization task, acquisition task, receiving and forwarding packets task, sending data package task, task of feeding dogs and standby task. The system block diagram of application is shown in Figure4. The processes of the system application in per round are as follows. After initializing the hardware resources, the tasks of collecting, sending, forwarding, standby and waking up are repeated circularly

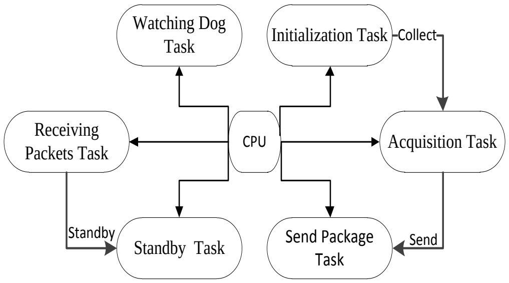


Figure 4. The application system diagram

IV. TEST AND RESULT ANALYSIS

A. *Node Hardware Performance Test*

The test of node hardware performance mainly refers to the test of low power performance. The working current in the system in different working modes is measured to test the performance of the hardware by using a high precision ammeter accessed to node circuit. After repeated testing, repeatedly recording the power consumption data of sensor nodes and calculating the average value, the test results are shown in Table 1. Results show that the sleep and wake-up mechanism can effectively extend the life cycle of nodes and achieve the requirements of low power consumption.

TABLE I. POWER TEST RESULTS OF SENSOR NODE

Test node	Standby	Data Collect	Data Send	Data Receive
Collectors/mA	0.024	32.32	26.25	21.95

B. *Network Packet Loss Rate Test*

In the networking test, 10 sensor nodes are divided into two clusters and a gateway node. The sensor nodes collect and send data once every 10min, and then enter the receiving mode after the data packet is sent. Finally, they will enter the sleep state after receiving the synchronization information from the gateway node. The test data of 14d are statistically analyzed and the results are shown in table 2. The results show that the system is stable and reliable, and the average packet loss rate of the network is 0.76%

TABLE II. PLR VALUES OF NETWORK

Cluster Num.	Node Num.	Send Packet Num.	Receive Packet Num.	Packet Loss Rate /%
1	01	2016	2005	0.54
1	02	2016	2001	0.74
1	03	2016	1999	0.84
1	04	2016	1995	1.04
1	05	2016	2004	0.59
2	01	2016	2002	0.69
2	02	2016	1997	0.94
2	03	2016	2001	0.74
2	04	2016	1995	1.04
2	05	2016	2006	0.49

V. CONCLUSIONS

To solve the problems occurred in on-line monitoring in large rice fields such as poor real-time and less system stability, large on-line monitoring fields, a node for rice field based on hybrid antenna is designed and its related performance test is carried out. The results of the test show that:

(1)By means of component selection, hardware circuit design and program design, a sensor node for rice field based on hybrid antenna is developed. The node can realize the stable real-time on-line monitoring towards the data of rice fields.

(2)Sensor nodes have low power consumption in the monitoring process.The current consumption of sensor nodes is 0.024mA under the sleeping mode, 32.32mA under the data collection, 26.25mA under data transmission and 21.95mA under the operating mode. In most of the time, the sensor nodes are in the standby sleep phase during the running time of the system so that the system can work stably in a long term.

(3)In the 14d networking of the labs, due to the close distance between nodes in this test, the average packet loss rate is 0.76%. Network packet loss rate relates closely with node deployment location, antenna height and other factors. During the process of practical monitoring, the network packet loss rate can be reduced by changing the transmission power of the wireless module, adjusting the height of the antenna, increasing the routing node and so on.

ACKNOWLEDGMENT

This research was financially supported by Science and Technology Planning Project of Guangdong Province, China (2014-17) and Science and Technology Program of Guangzhou, China （201605030013）

References

[1] Huizhe Chen, Defeng Zhu, The Rice Production and Ecosystem in the World, HYBRID RICE. 05(2003) 4-7.

[2] Zhenhuan Liu, Zhengguo Li, Pengqin Tang, et al. Spatial-temporal changes of rice area and production in China during 1980-2010, Acta Geographica Sinica. 68(5)(2013)680-693.

[3] Hongcheng Zhang, Jinlong Gong, Research Status and Development Discussion on High-Yielding Agronnmy of Mechanized Planting Rice in China, Scientia Agricultura Sinica. 07 (2014) 1273-1289.

[4] Feng Hong, Hongwei Chu, Zongke Jin, Review of Recent Progress on Wireless Sensor Network, Applications,Journal of Computer Research and Development..2010,47:81-87

[5] Xiaojun Qiao, Xin Zhang, Cheng Wang, Application of the wireless sensor networks in agriculture, Transactions of the CSAE.S2 (2005)232-234.

[6] Antai Han, Yong He, Zhiqiang Chen, et al. Design of Distributed Precision Irrigation Control System Based on Wireless Sensor Network for Tea Plantation, Transactions of the Chinese Society for Agricultural Machinery. 09 (2011)173-180.

[7] Jianqing Huang, Weixing Wang, Sheng Jiang, et al. Development and test of aquacultural water quality monitoring system based on wireless sensor network, Transactions of the Chinese Society of Agricultural Engineering. 04(2013) 183-190.

[8] Jin Hu, Hongpan Fan, Haihui Zhang, et al. Design of regulation system of light environment in greenhouse based on wireless sensor network, Transactions of the Chinese Society of Agricultural Engineering.04(2014) 160-167.

[9] Sheng Jiang, Weixing Wang, Daozong Sun, et al. Design of energy self-sufficient wireless sensor network node for orchard information acquisition, Transactions of the Chinese Society of Agricultural Engineering. 09(2012) 153-158.

[10] Heng Zhang, Dongyi Chen, Bing Liu, Research on the Orientation of Antenna in WSN, . Journal of University of Electronic Science and Technology of China, 2010(S1): 85-88.

[11] Jun Liu, Qian Sun, Shaohua Li, et al .Topology Control Algorithm Based on Directional Antennain Wireless AdHoc Networks, Journal of Northeastern University(Natural Science), 2012(09): 1257-1260.

[12] Guangsong Yang, Xu Geng. Engergy-efficient Route Finding Mechanism Based on Directional Antenna in Wireless Sensor Network, Computer Engineering, 2010(22): 91-93.

[13] Xiao Jun, Li Ke, Wang Jianhua.The Research of Communications and Route of Wireless Sensor Network[J]. Computer Knowledge and Technology, 2009. 2008, 2(18): 59-61.

[14] Xiaomin Li, Ying Zang, Xiwen Luo, et al. Design of WSN node with adaptive transmitting power for rice field, Transactions of the Chinese Society of Agricultural Engineering. 07(2014) 140-146.

[15] Baoxia Sun, Weixing Wang, Gang Lei, et al.Real-time Monitoring System for Paddy Environmental Information Based on Wireless Sensor Network, Transactions of the Chinese Society for Agricultural Machinery. 09(2014) 241-246.

Design of Real Time Monitoring System for Rural Drinking Water Based on Wireless Sensor Network

Jieping Yu
College of Electronic Engineering
South China Agricultural University
Guangzhou 510642, P.R.China
489398737@qq.com

Huili Yin
College of Electronic Engineering
South China Agricultural University
Guangzhou 510642, P.R.China
76576676@qq.com

Weixing Wang(Correspondence author),
Sheng JiangCollege of Electronic Engineering
South China Agricultural University
Guangdong Engineering Research Center for Monitoring
Agricultural Information
Guangzhou 510642, P.R.China
weixing@scau.edu.cn, jiangsheng@scau.edu.cn

Guohui Jiao, Zexin Lin
College of Electronic Engineering
South China Agricultural University
Guangzhou 510642, P.R.China
10353907779@qq.com, 995480993@qq.com

***Abstract*—With the continuous progress of rural construction, the problem of rural drinking water pollution is increasingly prominent. In view of water pollution, a design of rural drinking water monitoring system based on wireless sensor networks is proposed that nodes take STM32 as the core chip and WLK01L39 as well as its peripheral circuits are used as wireless communication modules and Beidou S1216 is used as GPS module to realize node localization. At the same time, the corresponding communication protocol and time synchronization algorithm are proposed. This paper have achieved automatic collection of water quality indicators, and uses GPRS network to achieve data upload. Experiments show the power consumption and data transmission performance of the system, and the packet loss rate is 6.2% when the communication distance of the system data is 150m in open area.**

Keywords-Water Pollution; Real-time monitoring; Wireless Sensor Network; Automatic acquisition; STM32

I. INTRODUCTION

With the rapid development of China's current economy, rural construction has been paid more attention to, and the process of rural modernization has been advancing. However, a series of environmental problems have followed[1-4].At present, there are three main problems in water quality monitoring in China. Firstly, the equipment in the water quality monitoring center is too old, and the allocation of large analytical instruments is inadequate, resulting in limited sampling and low monitoring frequency of the water quality monitoring center; Secongly, on-site monitoring capacity is lack; Last, water quality monitoring automation is not universal[5,6]. Therefore, a real-time monitoring system for water quality based on wireless sensor networks is proposed in this paper, which can remotely monitor water quality, ensure the real-time monitoring data, and effectively ensure the safety of rural drinking water[7，8].

II. OVERALL DESIGN OF WATER QUALITY REAL-TIME MONITORING SYSTEM

Indicators of water pollution include PH, dissolved oxygen, conductivity, turbidity, COD, BOD and so on[9].Water quality monitoring has many characteristics[10], such as large number of monitoring points, long monitoring time and complex monitoring environment. This system mainly consists of data acquisition node, convergence gateway and server monitoring center. Considering the characteristics of scattered points, the route forwarding node is also needed to avoid the loss of data acquisition due to the long transmission distance of nodes. The overall design of this system is shown in figure 1.

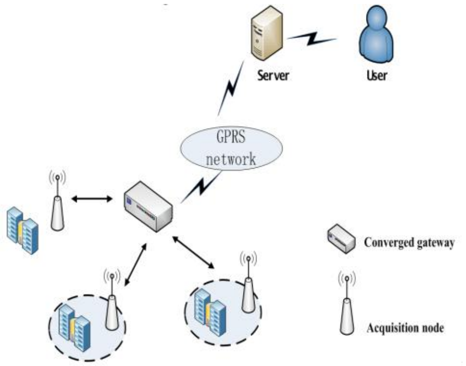


Figure 1. Overall design of system.

Data acquisition node is the foundation of the whole monitoring system. Acquisition nodes are distributed in monitoring waters, which are responsible for collecting water quality parameters and GPS location information which will be broadcast through the wireless module. The routing node is responsible for forwarding the received packets again, which can extend the data transmission distance between nodes, improve the stability of data transmission, and effectively reduce the packet loss rate. The converged gateway is responsible for connecting the server, obtaining the node number from the server, and the synchronization time. The converged gateway then uploads the received data from each node through GPRS, and then sends the synchronous packet to nodes to complete the synchronization of the whole wireless network.

The server is responsible for registering the number of nodes in the monitored area. When receiving the request of the gateway, it sends the matching node number and synchronization time to the gateway, and analyzes the data packet from the gateway. Finally, the relevant data is displayed in the monitoring center, which is convenient for users to query.

Different from traditional network systems[11], wireless sensor network have the characteristic of limited resources including limited power supply. The primary problem of wireless sensor network is the energy consumption[12].In order to prolong the life cycle of the whole network, the standby mechanism is adopted in the design. The data acquisition node and the gateway can complete the synchronization by using synchronous packets. After the end of the round, nodes and gateway will enter the standby mode to reduce the energy loss and prolong the life cycle of the whole network.

III. HARDWARE DESIGN

A. Design of Acquisition Node

According to the actual needs, the structure of wireless sensor network will vary, but it is usually composed of sensor unit, data processing unit, power supply and data transmission unit[13].

In this design, the acquisition node takes the STM32 microprocessor as the core, and collects the water quality parameters via sensors. The system uses Five Probe Sensor produced by Shanghai Jingji scientific instrument. The sensor can control the sensor to collect data by sending instructions, so it is convenient and efficient. If there is no location information corresponding to the data, the information is of no value[14],so the node is equipped with GPSS1216, which can be accurately positioned. At the same time, the data acquisition node has wireless communication module WLK01L39, which has low power consumption and long communication distance. On the premise of maximum transmit power, the communication distance of outdoor open area can reach 500m.The power module uses 12V large capacity lithium polymer battery to continuously supply power for all modules. The structure of the acquisition node is shown in figure 2.

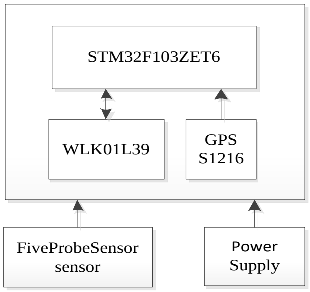


Figure 2. Design of acquisition node.

B. Design of Converged Gateway

The converged gateway is the center of the whole network. In addition to the GPS positioning module and wireless communication module, the convergence gateway is also equipped with GPRS module, which can connect to the server remotely, upload data and receive relevant information from the server. The converged gateway structure is shown in figure 3.

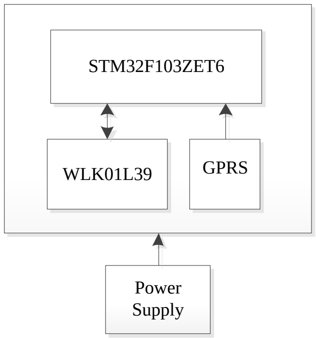


Figure 3. Design of converged gateway.

IV. SYSTEM SOFTWARE DESIGN

A. Software Design of Acquisition Node

The acquisition node is the terminal node of the whole network system. As shown in Fig. 4,the nodes performs GPS positioning, and then wait for the node matching packet from the gateway to determine the node network to which it belongs. The nodes wait for the synchronization packet to complete the time synchronization. After synchronization with the network, the water quality parameters are collected and encapsulated into packets, which are then sent out via a wireless module. When the standby time comes, the nodes enter the standby mode with the whole network, which can save the limited energy of nodes and prolong the working cycle of nodes.

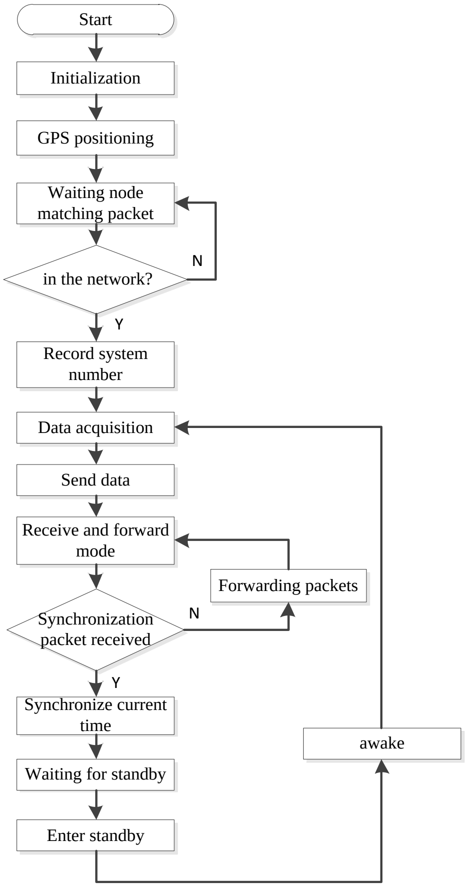


Figure 4. Software design of acquisition node.

B. *Software Design of Routing Node*

In wireless sensor networks, communication among nodes is easily affected by topography such as mountains, buildings and other obstacles[15], so nodes are easily separated from network. To ensure good transport performance, routing nodes are essential. The routing node is responsible for forwarding data packets from nodes and synchronous packets from the gateway. Routing forwarding nodes are shown in Fig. 5,which will be automatically set as forwarding mode after being awakened. When the standby time comes, the nodes enter the standby mode with the whole network, which can save the limited energy of nodes and prolong the working cycle of nodes.

C. *Software Design of Converged Gateway*

The converged gateway is shown in Fig. 6.The main tasks of the converged gateway including making and transmitting the node matching packets and the time synchronization packets and uploading the data from nodes. The gateway attempts to connect to the server after it is switched on. If the gateway successfully connects to the server, the server will return relevant information including node matching number, current time and a preset standby time. The gateway analyzes the received information from the gateway, makes node matching packets and synchronization packets, and then sends them out via wireless module.

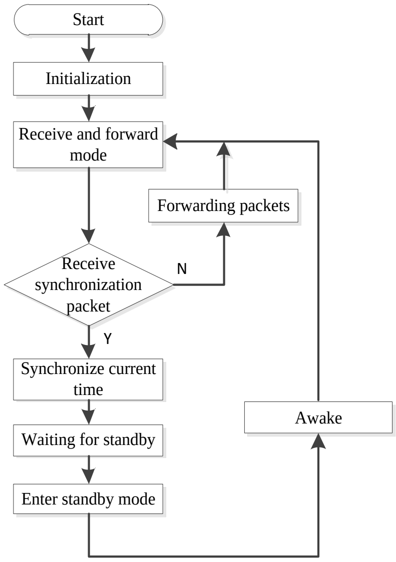


Figure 5. Software design of routing node.

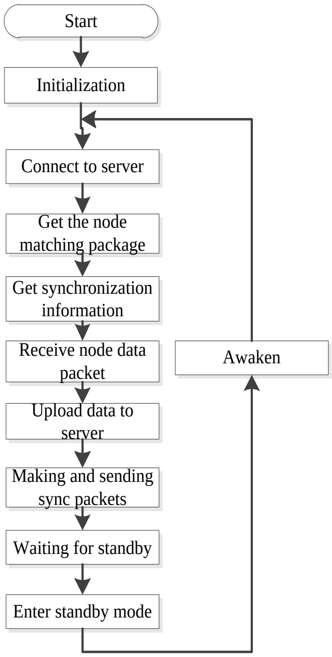


Figure 6. Software design of converged gateway.

V. EXPERIMENTAL RESULTS AND ANALYSIS

To test the power consumption of the system, an experiment was conducted at Hongze Lake in South China Agricultural University, which used a node and a gateway. Both the node and gateway are powered by 6800mAh batteries. The system continues to operate without replacing the battery. The node wakes up every 30 minutes and then works for about 40 seconds. The system started at 20:20 on December 10, 2016, and stopped uploading data at 16:40 on

December 21, 2016, running for about 11 days.In this process, a total of 520 packets were sent, 518 packets were actually received .Test result shows that the packet loss rate is0.38%. The actual test gateway is shown in Fig. 7

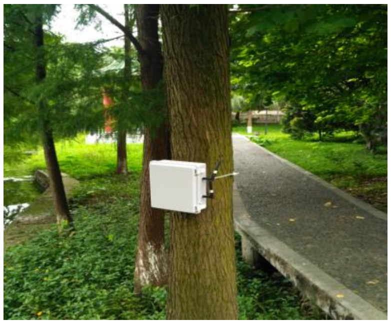
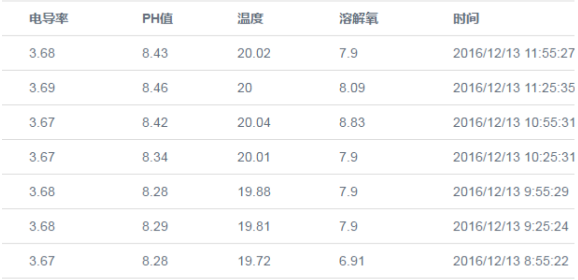


Figure 7. Monitoring site

Fig. 8. shows a partial record of the actual monitoring data.



电导率	PH值	温度	溶解氧	时间
3.68	8.43	20.02	7.9	2016/12/13 11:55:27
3.69	8.46	20	8.09	2016/12/13 11:25:35
3.67	8.42	20.04	8.83	2016/12/13 10:55:31
3.67	8.34	20.01	7.9	2016/12/13 10:25:31
3.68	8.28	19.88	7.9	2016/12/13 9:55:29
3.68	8.29	19.81	7.9	2016/12/13 9:25:24
3.67	8.28	19.72	6.91	2016/12/13 8:55:22

Figure 8. Recording of experimental data.

In order to test the stability of the transmission performance of the system, the test of packet loss rate of outdoor data transmission was carried out in Huashan stadium, South China Agricultural University, Guangdong Province. Without using the routing node, we selected the open area, sent a test data to the converged gateway every 5 seconds, and then counted the number of packets received. The test results are shown in table 1.

TABLE I. TRANSMISSION PERFORMANCE TEST

Distance	Number of data sent out	Number of data received	Packet loss rate
50m	500	500	0
100m	500	499	0.2%
110m	500	494	1.2%
130m	500	485	3%
150m	500	469	6.2%
200m	500	443	11.4%

VI. CONCLUSION

A real-time monitoring system based on wireless sensor networks is designed in this paper. In this paper, the whole structure of the system is described in detail, and the hardware design and software flow analysis are described, and the system operation and transmission performance were tested accordingly, which is of great significance in water quality monitoring. In the following design, it is necessary to do more in-depth research on the node energy consumption, networking methods and so on.

ACKNOWLEDGMENT

This research was financially supported by Science and Technology Planning Project of Guangdong Province, China (2014-17) and Science and Technology Program of Guangzhou, China （201605030013）

REFERENCES

[1] Ye Xiangbin, Chen Lihu, Hu Gang. Application of Wireless Sensor Networks in Environment Monitor[J]. Computer Measurement & Control, 2014, 1200(11): 1033-1035.

[2] Li Jinfeng, Liu Fengxi, Yang Zhouhua, Yang Xudong. Design of Water Quality Monitoring System based on Wireless Sensor Networks and GPRS[J]. Computer Measurement & Control, 2014, 22(12): 3887-3890.

[3] Zhang Guojie, Chen Kai, Yan Zhigang, Wang Wenhao. Water Enviroment Monitoring System based on Wireless Sensor Network[J]. Journal of Mechanical & Electrical Engineering, 2016, 33(3): 366-372.

[4] Luan Jian, Wang Qiang, He Xiaohui, Guan Fanglin, Bian Enjiang, He Jie.Analysis of Wireless Sensor Network Performance Evaluation[J]. Machine Building & Automation, 2015(3): 165-167.

[5] Jiang Liangzhong, Liu Qiao. Application Status and View of Auto Water Quality Supervision and Monitoring System in the Word[C]//Symposium on Application of environmental monitoring instruments and modern control technology in environmental control engineering.2003.

[6] Si Haifei, Yang Zhong, Wang Jun. Review on Research Status and Application of Wireless Sensor Networks[J]. Journal Mechanical & Electrical Engineering, 2011, 28(1): 16-20.

[7] Li Chengda, Zhang Jing, Gong Mingming.Overview of Wireless Sensor Networks and its Applications[J]. Journal of Chengdu Electromechanical College, 2007(3):10-14.

[8] Xiao Jun, Li Ke, Wang Jianhua. The Research of Communications and Route of Wireless Sensor Network[J]. Computer Knowledge and Technology, 2008, 2(18): 59-61.

[9] Legin A L, Ychkov E A, Vlasov Y G.Analytical Applications of Chalcogenide Glass Chemical Sensors in Environmental Monitoring and Process Control[J].Sensors and Actuators, 1995, 24(1/3):309-311.

[10] Wang Ji, Wang Xiaozhen, Ren Xiaoli, Shen Yuli. A Water Pollution Detective System Based on Wireless Sensor Network[J]. Journal of Guilin University of Electronic Technology,2009, 29(6): 247-250.

[11] Bi Ran, Li Jianzhong. Key Research Issues in Wireless Sensor Networks[J]. Intelligent Computer and Applications, 2014, 4(6):54-56.

[12] Zhao Jing, Pan Bin, Wang Jin, Tan Xiulan. Study on Energy Consumption and Strategies of Wireless Sensor Network, 2010, 43(10):87-88.

[13] Shi Junfeng, Zhong Xianxin, Chen Shuai, Shao Xiaoliang. Architecture and Characteristics of Wireless Sensor Network[J]. Journal of Chongqing University(Natural Science Edition), 2005, 28(2): 16-19.

[14] Zhu Huiyong. Research on Key Technology and Characteristics of Wireless Sensor Network[J]. Wireless Internet Technology, 2016(20): 12-14.

[15] Ying Jun. Features of Wireless Sensor Network and Energy Optimization Strategy[J]. Chongqing Technology Business University, 2008, 25(5): 537-540.