Conference Title

The Third International Conference on Electronics and Software Science (ICESS2017)

Conference Dates

July 31-August 2, 2017

Conference Venue

Takamatsu Sunport Hall Building, Takamatsu, Japan

ISBN

978-1-941968-42-0 ©2017 SDIWC

Published by

The Society of Digital Information and Wireless Communications (SDIWC)

Wilmington, New Castle, DE 19801, USA

www.sdiwc.net

Table of Contents

Six Kinds of Traffic Flow in Future Cities

Tomoya Yamamoto
Graduate School of Science and Engineering,
Hosei University, Kajino-cho 3-7-2,
Koganei-shi, Tokyo 184-8584
Email: tomoya.yamamoto.3a@stu.hosei.ac.jp

Takahiro Suzuki
Faculty of Science and Engineering,
Hosei University, Kajino-cho 3-7-2,
Koganei-shi, Tokyo 184-8584
Email: takahiro.suzuki.5q@stu.hosei.ac.jp

Isamu Shioya
Faculty of Science and Engineering,
Hosei University, Kajino-cho 3-7-2,
Koganei-shi, Tokyo 184-8584
Email: shioyai@hosei.ac.jp

***Abstract*—This paper presents a traffic flow model which considers self-driving in future cities. The model is based on a stochastic cellular automaton from the viewpoint of traffic resources in fundamental traffic diagrams. We suppose a future traffic system such as the updates of the model can be performed at every cells at the same time, so that every cars can move to forward at least one cell even if all the cells are occupied by cars. We introduce boundary effects and fluctuations as car-specific properties, and define six kinds of car running types based on the properties. Then, we discuss how six-running types are related to a traffic flow with respect to the variances of car speed in fundamental diagrams.**

***Index Terms*—Traffic flow, driver modeling, multi-agents, six-running types, and nonlinear features.**

I. Introduction

This paper presents a traffic flow model based on a stochastic cellular automaton which considers self-driving from the viewpoint of traffic resources in future cities. The future traffic system is that the updates of the model can be performed at every cells at the same time, so that every cars can move to forward at one cell even if all the cells are occupied by cars, where each cell is occupied at most one car. So, we can model to run the cars which are suitable to no time lags to move a car. We consider two unique features of effects as car-specific properties in running cars: boundary effects and fluctuations, and define six kinds of car running types based on the effects. The boundary effects are to accelerate car speed if there is the other car in behind, or to reduce car speed if there is other car within the range to move at one time step. The former is called a boundary effect in behind, and the later a boundary effect in front. The fluctuations are that the average moving speed of cars is shifted to either faster or slower randomly, so the fluctuations are distinct from a simple randomization. This paper discusses the features of the traffic flow which considers boundary effects and fluctuations, and we present the six kinds of car running types. Then, we discuss how six-running types are related to a traffic flow, and show that, in fundamental diagrams, each running type is quite different from others.

A smart city is an attractive concept to develop cities by using information technology and Internet of Thing (IoT). You may suppose that there are no traffic congestion in a smart city? Cities are chunks of human desires. So, a traffic accident may not happen, but a signal of desire is necessary. It might be quite difficult to prevent a traffic congestion, because a traffic resource, road, is finite and a lot of cars occupy the resources. In self-drivings in future cities, *autonomous* car moving is necessary to satisfy drivers desires. So, a car driving is sometimes sort of stressful work just as it is now. Suppose that you are now running a car in freeways. You may feel a higher stress to arrive within the scheduled time, when there is a car in front of your car or another car behind. Then, your car might accelerate/reduce the speed (a reaction to run cars, we call it a boundary effect). In another case, your car may be difficult to keep a car speed constant during running, because of uphills, downhills or others. Then, the speed of your car would be also accelerated/reduced (another running reaction, we call it a fluctuation effect). Our problem is to clarify how boundary effects and fluctuations are related to a traffic flow, where there are several running types as described later.

A variety of traffic flow models have been presented for analyzing a traffic flow and finding the solutions of the traffic congestion to escape. There are two kinds of traffic models to discuss a traffic flow: macro and micro models. A typical one of the macro model is based on Burgers equation from fluid mechanics[5]. The other micro model can be divided into two: Optimal Velocity (OV) model[1] and Cellular Automaton (CA) model [3]. OV model is presented for explaining traffic congestion based on differential equations. CA model [3] is successfully presented a fundamental model based on CA to reproduce traffic flow faithfully. But, it is sometimes difficult to explain the reason why the model is successful. This paper discusses how boundary effects and fluctuations are related to a traffic flow in fundamental diagrams, i.e. the number of car traffic vs. car density.

The works [4], [6] based on a probabilistic cellular automaton model employed the rules, which are similar to this paper, on a circle rather than our model discussed on a straight line: acceleration, slowing-down, randomization and car motion rules, and the model is trying to reproduce fairly faithfully a traffic flow in freeways. The work [2] which employed a slow-to-stop rule indicates that the results are close to the actual traffic flow.

Our original motivation of this paper is that a traffic flow is a stochastic autonomous moving multi-agent from the viewpoint of the efficient uses of traffic resources in moving multi-agent systems[7], each car is an agent in multi-agents, and each agent is arranged on the cells exclusively. Then, the fluctuations and boundary effects play important roles to get high resource utilization. This kinds of effects never go away. Because the

ISBN: 978-1-941968-42-0

drivers are not satisfied unless the cars autonomously move.

The main contribution of this paper is to see the traffic flow from two car feature aspects in the efficient uses of traffic resources: boundary and fluctuation effects. First: the boundary effects are an essential nature of a traffic flow, and are related to a transportation system. Our car update are performed by a fast car moving which is an ideal traffic flow without no time lag to move a car, i.e. the cell updates can be performed from right to left sequentially. This is suitable to run a car by self-drivings in future systems. Second: fluctuation effects are related to driving types which due to car specific features, so that the car speed is not only randomly changed but also the average car moving speed is randomly shifted to either slower or faster. We show how the boundary and fluctuation effects are related to traffic congestion to be happened in freeways under the fast car moving. They depend on their running types proposed six kinds in this paper.

This paper is organized as: the following section presents our future traffic flow model and discusses it. Our experimental results in Section III show how six-running types are related to a traffic flow with respect to the variances of car speed in fundamental diagrams. In the final Section IV, we conclude this paper.

II. Proposed Traffic Flow Model in Future Cities

We describe a stochastic traffic flow cell model whose cells are arranged on a one-way traffic resource, and the traffic is one lane (Figure 1 and 2). Each cell of the traffic is occupied by a car at most, and every cars stochastically run on the traffic resource from left to right.

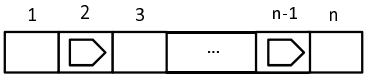


Fig. 1. Traffic resources, and cars run from left to right on a lane.

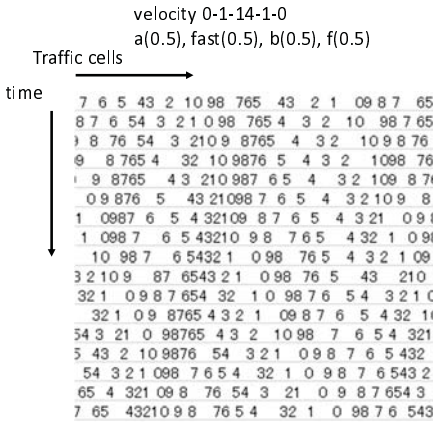


Fig. 2. The traffic resources in time domain.

The speed of each moving car is a stochastic process, and it depends on traffic rules, car performance, traffic flow and

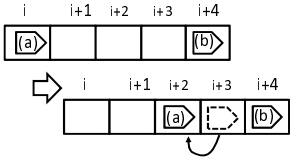


Fig. 3. Slower-in: Boundary effects in front.

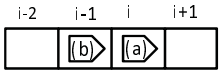


Fig. 4. Go-faster: Boundary effects behind.

car-specific characteristics. Suppose you are driving a car in freeways. We consider the *car-specific characteristics* as the following:

1) If there is a slower moving car ((b) in Figure 3) in front of your car, you ((a) in Figure 3) may slow down the speed more than necessaries in advance so that you might alleviate the stresses by slower-in and it's more safe. This is one of running types. On the other hand, if there is also another car ((b) in Figure 4) behind and there are no cars in ahead, the speed of your car ((a) in Figure 4) might be accelerated for avoiding from stresses behind, that is go-faster or faster-out. This is also one of running types. In this paper, the driver types that change the speed more than necessaries for avoiding collisions or alleviating higher stresses are said to be *boundary effects*. They depend on other cars stochastically.
2) As a special case of the boundary effects in above, if the cell in front of your car is unoccupied, but two cell in ahead and behind cell are occupied, your car either shorten one cell in front (see (a) in Figure 5) or not shorten, i.e. you either move a cell in ahead or remain there. We say that these car specific reactions are *minimal boundary effects* as described later by a.
3) Also, if there are small uphills or downhills in free-

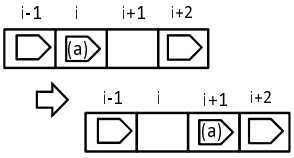


Fig. 5. Shortening one cell too close a car in front in minimum boundary effects.

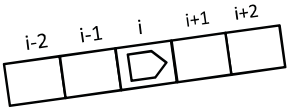


Fig. 6. An uphill in fluctuation effects.

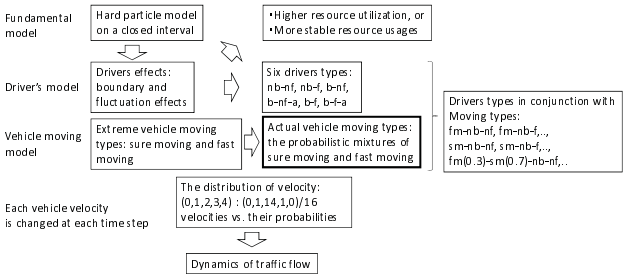


Fig. 7. Our proposed traffic flow model.

TABLE I
CAR RUNNING TYPES.

nb-nf	neither boundary nor fluctuation.
nb-f	allow fluctuation, but no boundary.
b-nf	allow boundary, but no fluctuation, no shorten, remain there, in minimum boundary.
b-nf-a	allow boundary, but no fluctuation, shorten one cell in minimum boundary.
b-f	allow both boundary and fluctuation, no shorten, remain there, in minimum boundary.
b-f-a	allow both boundary and fluctuation, shorten one cell in minimum boundary.

ways (Figure 6), you would be difficult to keep the car speed constant, and the speed of your car will be accelerated/slowed down in average speed. These also arise from some reasons including human factors or others, and do not depend on other cars. We say that their running types are *fluctuation effects* that the average moving speed of cars is shifted to either faster or slower randomly.

Now, how are you driving a car? We can consider two extreme car moving types in a traffic flow: "Fast moving" and "Sure moving". "Fast moving" is an ideal type of fast car moving in a traffic, so there are no time lags to move a car. Therefore, when the speed of all the cars are greater than or equal to 1, every car can move to forward at the same time by one step in CA, even if all the cells are occupied by car. That is, the updates in CA are performed from the most right cell to left cell, sequentially.

On the other hand, "Sure moving" is a type of a safe car moving at the next update step, so a car can move to forward, only if there are no cars in ahead at the previous update step. Actual car moving in freeways is the probabilistic mixture of two moving types: fast moving and sure moving. In this paper, we only consider the fast moving, and discuss the relationships between car density and traffic flow under six car running types, also described in later, of running reactions.

Our model [7] is based on a hard particle model on closes intervals, and the goal is to achieve higher resource utilization or stable multi-agent behavior. There are two magics for getting higher resource utilization in multi-agents: a fluctuation (a shake) and a boundary effect. The later depends on their positions. The two are just psychological effects in car driving. At the same time, they lead to effective use of traffic resources (Figure 7).

In the following section, our experiments show how the six kinds of running types are related to the traffic flow in fundamental diagrams.

III. EXPERIMENTS

We consider six kins of car running types based on the running reactions in Section II, shows them in Table I, and assign the symbols nb-nf,.., and b-f-a to their running types.

We prepare a one-way traffic consisting cells, and the traffic is one lane. Every car moves on the traffic cells from left to right shown in Figure 1 rather than on circles. The number of the traffic cells is $2,000$ on the straight. The moving type which is considered is only fast moving, so we perform the updates of the traffic cells from right to left, reverse, at each moving step, and provide $30,000$ cars. Their cars run $30,000$ steps for each exam in our experiments. At each updating step, a car enters into the traffic only if the leftmost traffic cell is empty. Otherwise, the car has to wait to enter into the traffic until left most cell becomes empty. This means that we examine the maximum traffic flow for every driving type in fast moving on one way lane without signal lights, freeways.

Assume that each car speed v is either 0, 1, 2, 3 or 4 in the cases without running reactions, and the average value is 2 if there are no cars in ahead. Each car speed randomly changes at every moving step. We use Mersenne twister as a random number generator.

The boundary effects are implemented as every car speed v either increase 1 in average if the behind cell is occupied and two more cells in ahead are unoccupied, or decrease 1 in average if v more cells in ahead are unoccupied and the car moves to the behind of car in ahead with a slowdown in speed. Shortening one cell in the minimal boundary effect (Figure 5) is considered if there only exists an empty cell in front and the speed is greater than or equal to 1. With even probability, the fluctuation effects either increase the speed 1 or decrease the speed 1 in average.

We examine the car density and the traffic flow by varying the distributions of the speed 0, 1, 2, 3 and 4 correspond to the probabilities 0, $\frac{1}{16}$, $\frac{14}{16}$, $\frac{1}{16}$ and 0, respectively. Then, we observe the car density on the cells between 501th and 1,500th, and the traffic flow is the number of cars to pass on the 2,000th cell between 20,001th and 30,000th steps. We performed them 10 times for each running type, and obtained the traffic flow versus car density relationships, the fundamental diagrams, shown in Figure 8-14. Our problems are how six-running types are related to a traffic flow with respect to the variances of car speed, and we show them in the fundamental diagrams of traffic flow.

Figure 9 presents the fundamental diagram of nb-nf by changing the variation of car speed. 00-01-14-01-00 (the speed 0, 1, 2, 3 and 4, respectively) in the figure indicates the car speed distribution ratios which correspond the probabilities with $\frac{0}{16}$, $\frac{1}{16}$, $\frac{14}{16}$, $\frac{1}{16}$ and $\frac{0}{16}$. It is quit a sharp convex

distribution. That is, every car speed is almost the same, but there is a slightly different car. On the other hand, 05-03-00-03-05 in the figure indicates the car speed distribution ratios which correspond the probabilities with $\frac{5}{16}$, $\frac{3}{16}$, $\frac{0}{16}$, $\frac{3}{16}$ and $\frac{5}{16}$. It is a concave distribution. That is, there are a mixture of extremely slow cars and cars trying to run fast. When gradually changing from a concave speed distribution to a convex one, the number of cars and car density in the fundamental diagram are increased together, and finally turn to left upside.

Figure 10 shows the fundamental diagram of nb-f by changing the variation of car speed. It is a concave distribution. When gradually changing from a concave speed distribution to a convex one, the number of cars and car density in the fundamental diagram are increased together.

Figure 11 shows the fundamental diagram of b-nf by changing the variation of car speed. It is a concave distribution. When gradually changing from a concave speed distribution to a convex one, the number of cars and car density in the fundamental diagram are increased together, and finally skip up to top.

Figure 12 show the fundamental diagram of b-f by changing the variation of car speed. It is a concave distribution. When gradually changing from a concave speed distribution to a convex one, the number of cars and car density in the fundamental diagram are slowly increased together.

Figure 13 shows the fundamental diagram of nb-nf-a by changing the variation of car speed. It is a concave distribution. When gradually changing from a concave speed distribution to a convex one, the number of cars and car density in the fundamental diagram go from the lower right to the upper left.

Figure 14 shows the fundamental diagram of b-f-a by changing the variation of car speed. It is a concave distribution. When gradually changing from a concave speed distribution to a convex one, the number of cars and car density in the fundamental diagram are increased together, and finally skip up to top.

Figure 9-14 shows that boundary effects are closely related to fluctuation effects, so that both boundary and fluctuation effects are of essential to achieve efficient utilization of traffic resources. Our results show that we feel stresses in driving and take some actions as running types so that the car density in traffic flow is improved drastically.

IV. Conclusions

We have examined how running types are related to the maximum traffic flow in an ideal car fast moving. Then, we have considered six running types to run cars: nb-nf, nb-f, b-nf, n-nf-a, b-f and b-f-a, so that the traffic flow depends on the six running types. And, we discuss how six-running types are related to a traffic flow with respect to the variances of car speed in fundamental diagrams.

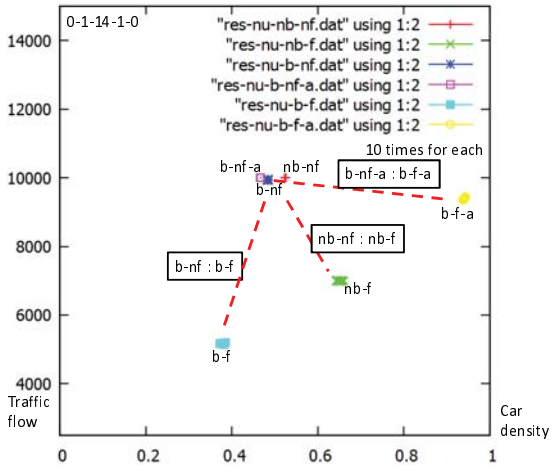


Fig. 8. Traffic flow versus car density relationships among car driving types: the probabilities for the speed 0, 1, 2, 3 and 4 correspond to 0, $\frac{1}{16}$, $\frac{14}{16}$, $\frac{1}{16}$ and 0, respectively.

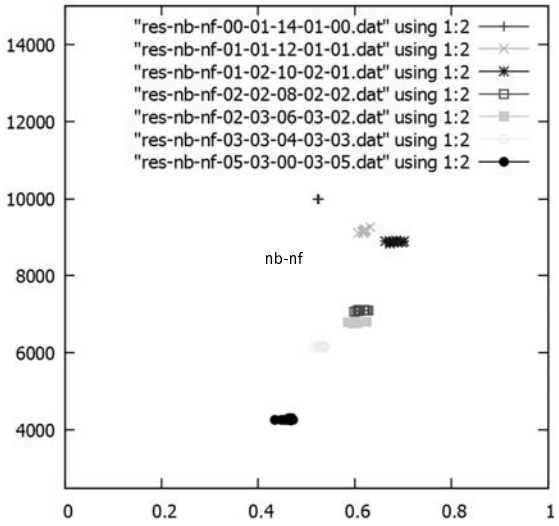


Fig. 9. Traffic flow versus car density relationship of nb-nf by changing the variances of car speed: the probability ratios for the speed 0, 1, 2, 3 and 4 are illustrated at the upper side in the figure.

References

[1] M. Bando, K. Hasebe, K. Nakanishi, A. Nakayama, A. Shibata, and Y. Sugiyama. Phenomenological study of dynamical model of traffic flow. *J. Phys. I France*, 5:1389–1399, 1995.

[2] A. Clarridge and K. Salomaa. Analysis of a cellular automaton model for car traffic with a slow-to-stop rule. *Lecture Notes in Computer Science*, 5642:44–53, 2009.

[3] B. Derrida, M.R. Evans, V. Hakim, and V. Pasquie. Exact solution of a 1D asymmetric exclusion model using a matrix formulation. *J. Phys. A: Math. Gen.*, 26:278–287, 1993.

[4] K. Nagel and M. Schreckenberg. A cellular automaton model for freeway traffic. *J. Phys. I France*, 2:2221–2229, 1992.

[5] F.T.M. Nieuwstadt and J.A. Steketee. Selected papers of J.M. Burgers. *Springer*, 1995.

[6] A. Schadschneider and M. Schreckenberg. Cellular automaton models and traffic flow. *J. Phys. A: Math Gen.*, 26:L679–L683, 1993.

[7] I. Shioya. Resource utilization of stochastic mobile multi-agents. *IEEE PACRIM*, pages 137–141, 2013.

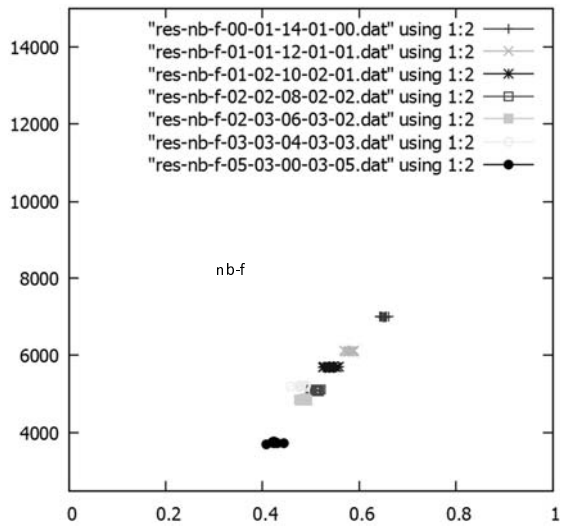


Fig. 10. Traffic flow versus car density relationship of nb-f by changing the variances of car speed: the probability ratios for the speed 0, 1, 2, 3 and 4 are illustrated at the upper side in the figure.

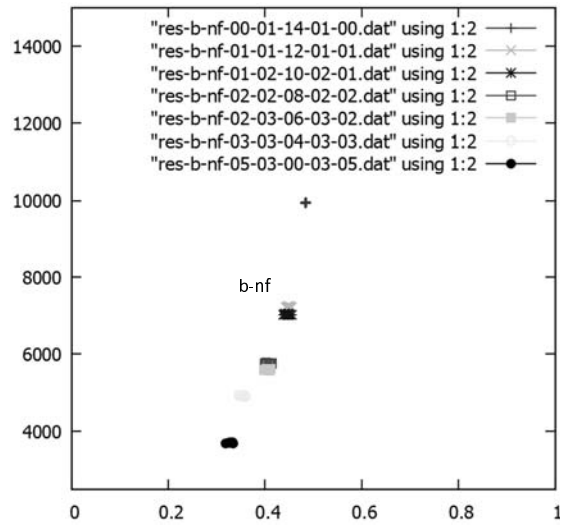


Fig. 11. Traffic flow versus car density relationship of b-nf by changing the variances of car speed: the probability ratios for the speed 0, 1, 2, 3 and 4 are illustrated at the upper side in the figure.

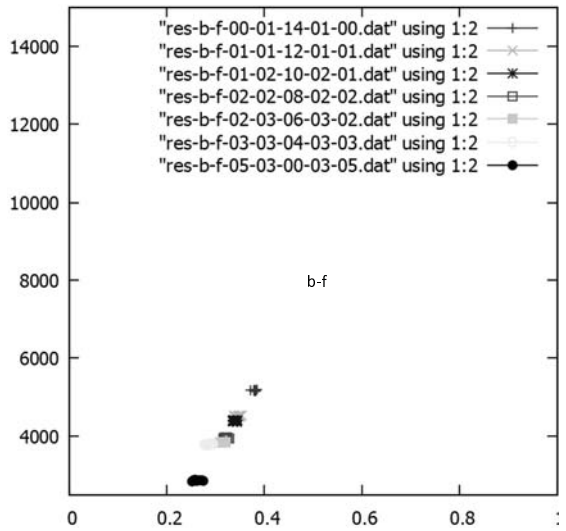


Fig. 12. Traffic flow versus car density relationship of b-f by changing the variances of car speed: the probability ratios for the speed 0, 1, 2, 3 and 4 are illustrated at the upper side in the figure.

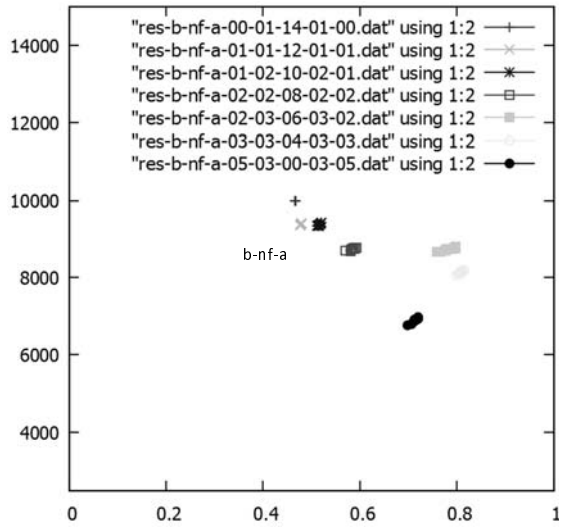


Fig. 13. Traffic flow versus car density relationship of b-nf-a by changing the variances of car speed: the probability ratios for the speed 0, 1, 2, 3 and 4 are illustrated at the upper side in the figure.

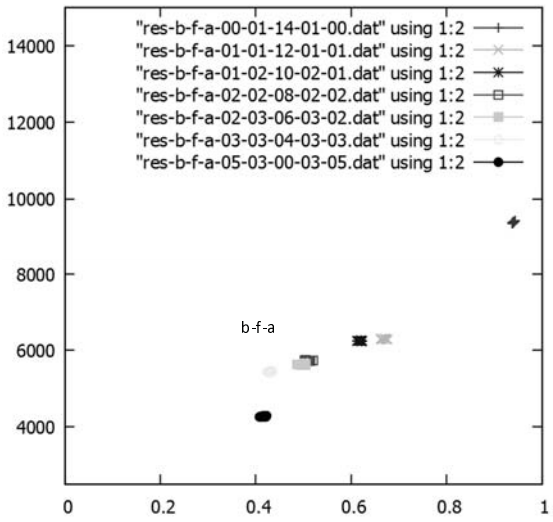


Fig. 14. Traffic flow versus car density relationship of b-f-a by changing the variances of car speed: the probability ratios for the speed 0, 1, 2, 3 and 4 are illustrated at the upper side in the figure.

Design, Implementation and Trial Evaluation of CPU Simulator to Visualize Register-transfer level Micro-Operation

Shinya Hara* Yoshiro Imai Koji Kagawa Kazuaki Ando
Rihito Yaegashi Keizo Saisho Kyosuke Takahashi Hitoshi Inomo
Naka Gotoda Toshihiro Hayashi Hiroyuki Tominaga Tomohiko Takagi
Graduate School of Engineering, Kagawa University
2217-20 Hayashi-cho, Takamatsu, Kagawa pref., Japan
s17g477@stu.kagawa-u.ac.jp*,
{imai,kagawa,ando,rihito,sai,k_taka,inomo,gotoda,hayashi,tominaga,takagi}@eng.kagawa-u.ac.jp

ABSTRACT

This paper proposes a new educational tool for Computer Architecture, which can provide simulation of assembly program code (instead of machine language), demonstration of several kinds of sample programs and visualization of register-transfer-level structure/behavior, namely micro-operation. Our educational tool for CPU simulation has been designed and implemented in Javascript language as Web service. Its users select simulation modes by micro step, by machine cycle and by automatic repetition of such cycles. So they can learn how a computer works graphically, recognize inner structure of CPU and understand micro-operation based behavior of CPU. Our Simulator has been also evaluated through some kinds of questionnaires by users/learners in classroom lectures. It is confirmed that the simulator has been very useful and effective to learn Computer Architecture and behavior/organization of CPU by means of its application.

KEYWORDS

Educational Visualization, Computer Simulation at Register-transfer level, e-Learning.

1 INTRODUCTION

It is very much important for students of universities and higher education to understand structures and behaviors of computer precisely not only during their school days but also for the sake of their future. There have been any useful efforts for researchers and professors to realize effective educational tools for the above students in order to provide fruitful results of learning. Such educational tools include simulators and e-Learning systems[1][2][3][4]. They have been evaluated by users, teachers and researchers suitably and significantly.

We have developed a new CPU simulator for learners to recognize internal structure and register-transfer level micro-operation of CPU of computer graphically and concretely. Our CPU simulator can visualize machin-cycle level behavior as well as micro-operation level one. It has been designed and implemented with JavaScript for its smart execution on the major browsers of PCs as well as tablets/smart phones. The former have been used at home and in the classroom lectures, while the latter becomes more and more popular in our daily lives. Due to these factors and its facilities, the CPU simulator has been able to play an important role to educate students to understand how a computer works from the viewpoint of micro-operation.

This paper describes design, implementation and trial evaluation of our CPU Simulator which can visualize register-transfer level micro-operation and show how a computer works graphically. The paper illustrates system configuration of our CPU simulator in the next section, demonstrates how to use the simulator in the third section, reports its evaluation through some questionnaires, and finally summarizes the conclusion in the last section.

2 SYSTEM CONFIGURATION

This section describes user interface of the simulator, four types of simulation modes, and

representative facilities from the simulator.

2.1 User Interface

Our simulator can run on the major browsers of PCs and tablets/smart phones because of its implementation by JavaScript. It is important for learners to handle the simulator smoothly and correctly in order to understand how a computer works graphically. User interface has been designed and implemented carefully for learners to recognize inner structures and behaviors of CPU more clearly. Figure1 shows user interface (namely UI) of our CPU simulator.

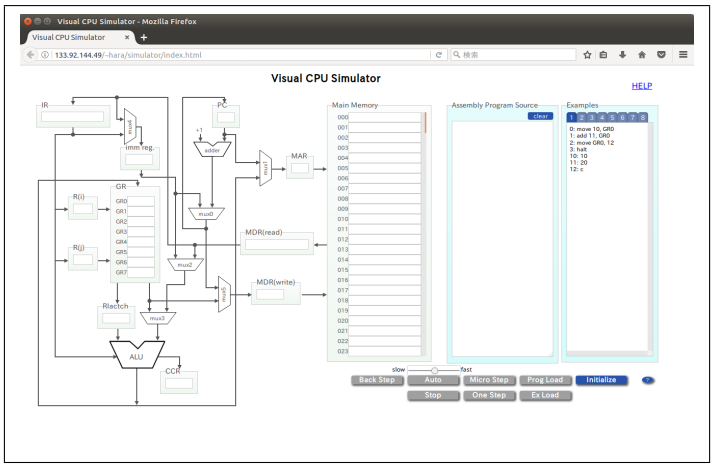


Figure 1. User Interface of our CPU Simulator on UBUNTU Linux

The left side of UI in Fig1 shows detail of CPU and main memory. The right side of UI also shows notepad and list of sample programs, the former can provide note space for learners to write their programs (written in assembly codes) and the latter expresses a list of the registered sample programs learners can select to execute by simulator. There are a group of buttons for learners to manipulate our CPU simulator very easily. Those buttons include system initializing, program loading, program executing and so on.

2.2 Four Types of Simulation Modes

Our CPU simulator equips and presents four useful modes of execution of loaded programs. They are summarized as follows;

1. machine-cycle level (one step execution):
   Our CPU simulator can visualize inner behavior of CPU, namely instruction fetch, instruction decode and instruction execute. With one step execution, learners can understand mechanism of CPU which performs instruction fetch, instruction decode and instruction execute graphically.

2. repetition of machine-cycle level until stop by users (autonomous execution):
   By repetition of machine-cycle behavior, the simulator provides autonomous execution for a series of assembly codes. With this mode, learners can confirm whether focused program or a series of assembly codes (namely a part of program) runs/executes correctly or not.

3. micro-operation level:
   Especially, our CPU simulator can visualize register-transfer level of CPU behavior, for example, instruction fetch has been decomposed to the following micro-operation such as “transfer the content of Program Counter(PC) into Memory Address Register(MAR)”, “read the relevant data of memory addressed by MAR”, and “transfer such data into Memory Data Register for Reading(MDR(Read))”. With micro-operation behavior, learners can understand suitably inner structures of CPU and register-transfer level of CPU behavior.

4. backward execution of micro-operation level:
   And more especially, our CPU simulator can utilize one step backward execution of micro-operation. With the relevant operation, not only learners but also teachers can investigate the execution of CPU in the forward direction as well as in the backward one in order to check and confirm their program.

With these four manipulation modes, users can utilize our CPU simulator efficiently and effectively.

ISBN: 978-1-941968-42-0 ©2017 SDIWC

2.3 System Facilities

Our CPU simulator also prepares some useful facilities, for example, as follows;

- explanation of usage:
  Intro.JS has been employed to realize smart explanation of usage of our CPU simulator. It will bring good performance of usage to our users. The same approach had been employed in our previous simulators and well-evaluated by the relevant users [4][5].

- speed control of simulation:
  In the mode of repetition of machine-cycle level, our users can select speed of simulating by means of speed control slider shown in Figure2. With this facility, our simulator realizes suitable speed of CPU simulating for the focused program in order to understand how a computer works graphically.

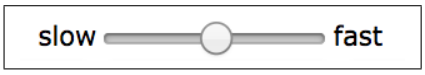


Figure 2. variable speed control of simulation

3 PERFORMANCE OF SIMULATOR

This section demonstrates system performance of our simulator for user servies and introduces how to utilize it in order for learners, user in the other hand, to understand how a computer works graphically.

3.1 System Performance

After our CPU simulator is invoked on the major browser(s) such as Mozilla FireFox, Google Chrome, Microsoft Edge/IE and Mac Safari on user's PC, it will provide the following typical manipulation by user, namely

1. Preparation of program:
   Users can select two modes for preparation of program to be executed by our CPU simulator. One mode is for users to write their assembly codes directly in the space called "Assembly Program Source" in Figure1. Another is for them to choose an example of programs from the list called "Examples" at the right hand of its UI in Figure1. The relevant program, namely assembly code, will have been ported into the memory of CPU simulator called "Main Memory" at the middle of UI in Figure1.

2. Initialization of CPU simulator:
   Before executing the loaded program in the main memory, it is necessary to accomplish a suitable initialization of the elements of CPU, such as General-purpose registers (GR), Program Counter and so on. Our simulator prepares a button to do initialization of CPU simulator not only for some registers of CPU but also all the contents of main memory. So users had better click such an initialization button at first and click other buttons such as program loading secondarily.

3. Program loading:
   Program loading is also necessary before program executing. Such a loading has two different options, namely "Prog Load" to transfer into main memory from the contents of the area called "Assembly Program Source" and "Ex Load" to transfer into main memory directly from the contents of the selected example of the list called "Examples". Users can instruct our simulator to perform program executing after the adequate contents of main memory had been occupied with assembly code from "Assembly Program Source" or one "Examples".

4. Program executing:
   After program loading is finished, program executing can be carried out by means of choosing one of the four powerful modes for execution of simulating. The detail of those manipulation will be explained in subsection2.2. In the mode of "autonomous execution", users can adjust simulating speed as was explained at the latest part of subsection2.3.

3.2 Simulation at Register-Transfer Level

In our Simulator, micro-operations are designed to include 10 numbers of micro steps, so that learners can recognize visually that one machine-cycle has been realized by means of multiple register-transfer level micro-operations in CPU through usage of our simulator.

In this case, we explain how one machine-cycle, "move 10, GR0" which instructs the content of memory addressed by 10 to be transferred into General-purpose register GR0, is realized with register-transfer level micro-operations. We will demonstrate our CPU simulator to visualize these relevant register-transfer level micro-operations below. After initializing and program loading, buttons of "Micro Step" and other executing modes are available. Please check Figure1. In such a case, buttons of "Micro Step" and other executing modes were not available before initializing and program loading, because those buttons had gray-colored backs.

After Instruction Fetching and Decoding, Address Calculation and Data Calculation have to be carried out. Figure3 shows Address Calculation to access the data of memory addressed by MAR whose content is "10" in yellow back. This figure demonstrates register-transfer from "imm reg." in light blue back (means source) to "MAR" in yellow back (means destination) through blue line. Figure4 shows Data Calculation/Strorage to obtain the data of "GR0"

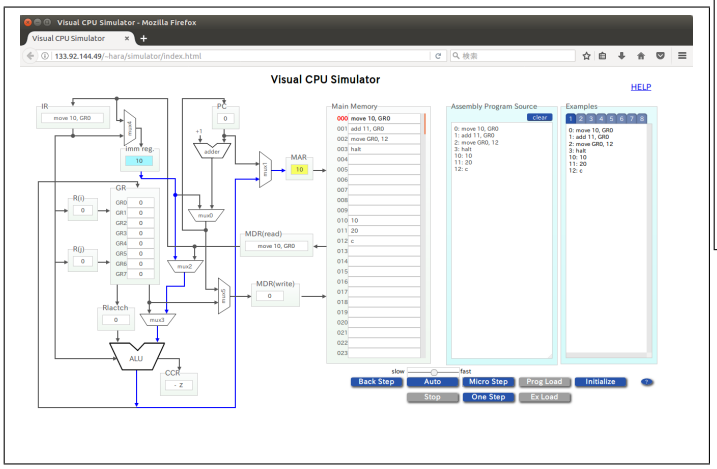


Figure 3. Register-Transfer for Address Calculation.

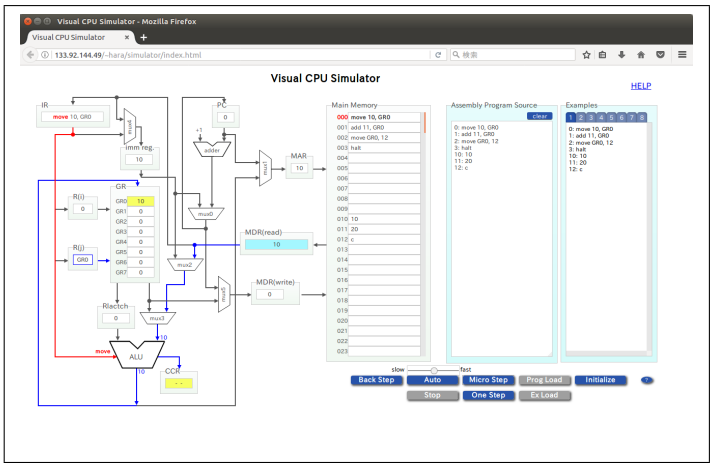


Figure 4. Register-Transfer for Data Calculation/Storage.

from "MDR(read)". This figure demonstrated register-transfer from "MDR(read)" in light blue back to "GR0" in yellow back through blue line with reflect for Condition Code Register(CCR). In this case, red line plays a role to transfer control signals generated by instruction decoding. After Data Calculation/Storage, Program Counter(PC) must be updated/modified, Figure5 shows PC to be incremented by "+1" with usage of dedicated adder through blue line. And PC also plays destination so that PC has yellow back.

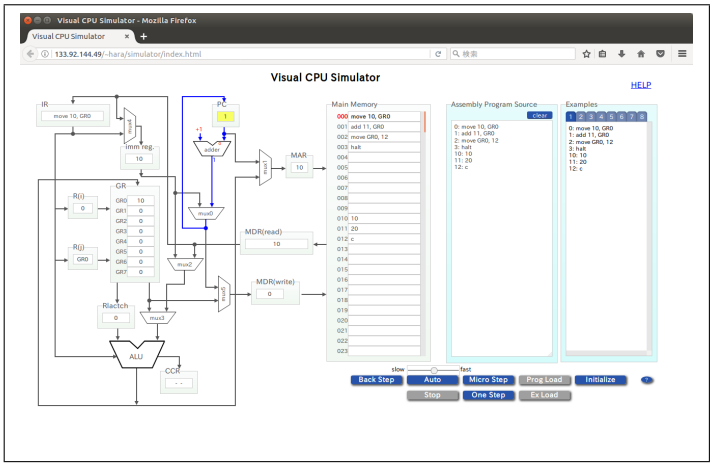


Figure 5. Update of Program Counter

Our CPU simulator can visualize Register-transfer level behavior of inner CPU and also demonstrate how CPU executes loaded program by means of micro-operations.

4 EVALUATION

This section explains the contents of questionnaire and its results as evaluation of our simulator through users in the classroom lectures.

4.1 Contents of Questionnaire

After usage of our simulator in the practical classroom lecture, we had carried out the following seven numbers of questionnaire for our learners who were the first year students of our departments. Learners were asked to choose five options from Grade5(Excellent) to Grade1(Poor) for below each question.

(1) Do you think that our simulator can help you to understand data flow inside of computer ?

(2) Do you think that our simulator can help you to understand what kinds of micro-operations organize machine instruction ?

(3) Do you think that multiple execution modes of our simulator can help you to learn how a computer works ?

(4) Are the display facilities such as color assignment and layout useful for you to manipulate our simulator ?

(5) Is change of color for registers and connection lines useful for you to understand data flow inside of CPU ?

(6) Is simulator manipulation easy for you ?

(7) Does usage of our simulator bring motivation for learning to you ?

The aim of the above questions are to confirm whether characteristics and merits of our CPU simulator can provide user's benefits and contribution to learner's understanding. Such a questionnaire was performed as voluntary base and only 34 students kindly replied their answers about the relevant questions, while 40 students participated in our classroom lecture.

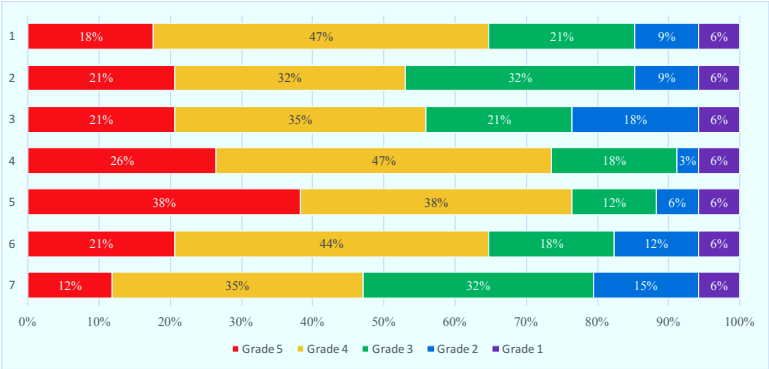


Figure 6. BarChart graph for the result of questionnaire: Red for Grade5 · · · Purple for Grade1

4.2 Results of Questionnaire

Results of the above questionnaire is summarized in bar-chart graph of the Figure6 as follows. Evaluation of our simulator based on the above results for questionnaire has been summarized as follows;

- Because of the results of questions (1) and (5), namely approximately more than 65% of answers are agreeable for the relevant questions, it may be confirmed that our simulator can achieve understanding data flow inside of computer/CPU.

- However, because of the results of questions (3) and (6), namely approximately less than 60% of answers only are agreeable for the relevant questions, some manipulating and supporting facilities cannot obtain users' good evaluation.

5 CONCLUSION

We have developed CPU Simulator written in Javascript to demonstrate how a computer works graphically. It can execute in almost all the major browsers not only of PCs but also of Tablets/Smartphones. One of the useful characteristics of our CPU simulator is to visualize Inside structure and behavior of a CPU through the Register-transfer level micro-operations. With its four executing modes, users can manipulate our CPU simulator towards maximal performance and learners can understand how a computer processes its pro-

ISBN: 978-1-941968-42-0 ©2017 SDIWC

gram codes at the level of micro-operations effectively and efficiently.

We have carried out questionnaire with eight items and evaluated our system through the results of the questionnaire. From the result of questionnaire, we think it is confirmed that our simulator realizes good users' learning and understanding how a computer work graphically. Approximately more than 65% of the answers about the relevant characteristics and merits can be considered to be good and agreeable by our users. At the same time, approximately less than 60% of the answers for some manipulating and supporting facilities can be considered to be agreeable by our learners, so that we must improve such facilities near future.

REFERENCES

[1] M. Grigoriadou, et al., "A Web-based Educational Environment for Teaching the Computer Cache Memory," IEEE Transactions on Education, Vol.49, No.1, pp.147–156, May 2016.

[2] Y. Imai, et al., "Application of a Visual Computer Simulator into Collaborative Learning," Journal of Computing and Information Technology, Vol.14, No.4, pp.267–273, Dec. 2006.

[3] C. Kawanishi, et al., "Development and Evaluation of Learner-centric Graphical Educational Tool for Network Study," Proc. of 2013 International Conference on Humanized Systems (ICHS2013@Takamatsu), pp.108–113, Sept. 2013.

[4] K. Higashikakiuchi, et al., "Design, Implementation and Trial Evaluation of a Visual Computer Simulator by JavaScript," Proc. of The Second International Conference on Electronics and Software Science (ICESS2016@ Takamatsu), pp.158–164, Nov. 2016.

[5] K. Nishiyama, et al., "Design and Implementation of a Scheduling Algorithms Visualizer using JavaScript," Proc. of The Second International Conference on Electronics and Software Science (ICESS2016@ Takamatsu), pp.175–181, Nov. 2016.

Multilingual Improvement of an e-Learning System for Packet Routing Visualization

Lorkan Sauvion* Valentin Messias* Chiaki Kawanishi Yoshiro Imai
Koji Kagawa Kazuaki Ando Rihito Yaegashi Naka Gotoda
Kyosuke Takahashi Hitoshi Inomo Tomohiko Takagi Hiroyuki Tominaga
Toshihiro Hayashi Shunsuke Doi Shinya Hara Tetsuo Hattori
*Engineering Degree in Computer Science, ESIEE Paris,
*Cité Descartes, 2 Boulevard Blaise Pascal, 93160 Noisy-le-Grand, France
*{lorkan.sauvion, valentin.messias}@edu.esiee.fr
Graduate School of Engineering, Kgawa University
2217-20 Hayashi-cho, Takamatsu, Kagawa pref., Japan
{imai,kagawa,ando,rihito,gotoda,k_taka,inomo,takagi,tominaga,hayashi,hattori}@eng.kagawa-u.ac.jp

ABSTRACT

Kagawa university had already developed an e-Learning system with self-learning oriented facility of application program for the sake of studying network/communication/packet-controlling. Important facilities of that system were to provide network topology design with routers and hosts and then to demonstrate the relevant behavior based on packet routing with animation. This time, we improve that system into a world-wide available e-Learning system, which can equip multilingual supports and explanations not only for Japanese learners but also for French/English/Portuguese spoken learners. Thanks to these multilingual supports and explanations, a newly improved system will bring a very useful educational environment to learners of the world and will let them manipulate our newly modified e-Learning system for network study effectively and efficiently. We will try to evaluate our system through our performing questionnaire and receiving comments from user of the world near future.

KEYWORDS

e-Learning tool for network education, Visualization of network topology and IP routing, Multilingual support.

1 INTRODUCTION

As Internet becomes over ground and most effective for our lives, in higher education, especially information engineering, network and communication are ones of the most important and indispensable subjects for almost all the students to learn in a relatively short period. There are many trials to teach and educate some theoretical and practical understanding of network and communication[1][2]. In fact, however, network and communication, particularly a viewpoint of Internet, have a very huge scope and include a lot of themes to understand at university as well as in the same higher education. Therefore, Kagawa university developed a limited scope and self-learning oriented facility of application program as an e-Learning system for network/communication study[3]. Our key topics are as follows: Web-based e-Learning system, Focusing scheme of IP(Internet Protocol), Designing/Easy-drawing network topology, Understanding network behavior through animation of packet transmission, Illustration of routing mechanism, and so on. These trials were very valuable and useful for Japanese learners only because their description/explanation language was Japanese[4]. If multilingual explanation had been available for them, such facilities and/or educational tool would be more and more useful and attractive for foreign learners. Since 2012, ESIEE Paris[1] and Kagawa university Japan have been exchanging International Internship Program so that we have applied for this program to visit Kagawa university Japan

[1]previously named École Supérieure d'Ingénieurs en Électrotechnique et Électronique

and to study at Prof. Kagawa-Imai laboratories for world-wide research, namely our "international project", about Domain of Information Science. We are trying to redesign/modify well-evaluated e-Learning system and improve it with providing world-wide availability in order for foreign learners as well as Japanese students to understand the relevant system effectively and efficiently.

This paper introduces short briefing about characteristics of the e-Learning system in the next section. It illustrates multilingual improvement for the relevant system for French/English/Portugese spoken learners to manipulate this system based on multilingual supports and then demonstrates some advantages of multilingual e-Learning system with practical examples. Finally, the paper concludes our summaries about this project for more precise evaluation near future.

2 EDUCATIONAL TOOL FOR NETWORK STRUCTURES AND BEHAVIORS

The first half of this section describes outline and user interface of the e-Learning system for network study, and second one introduces modification of JavaScript source codes for multilingual facilities.

2.1 Outline and User Interface of System

It is important to recognize network topology and data flow between nodes in order to understand network-related subjects in universities such as a concept of 'Packet', tasks of each component, namely nodes and router, and so on. At the view point of IP-based network structure and behavior, 'IP Routing' can play a very important role to specify the network structure and define packet transfer from one to another. People say that IP routing seems to be one of the most useful layers to learn the computer network. That is very much centered in the network-related subjects. And it is one of the reasonable approaches for learners to begin studying design and recognition of structure and behavior of Network. Kagawa university's approach had focused these IP routing mechanism and structure and behavior of network, because their aims realize a shorter period to understand basic IP routing by means of suitable educational tool for network study. Its outline is as follows;

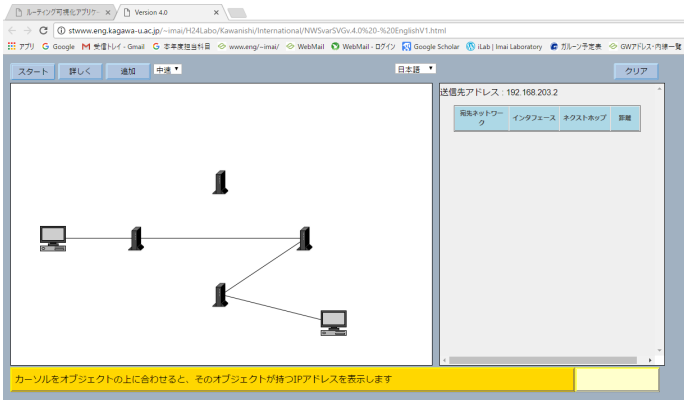


Figure 1. User interface for Japanese users(riginal).

- allocate nodes/routers
- connect them and organize network
- assign source/dest. of packet routing
- visualize IP routing through animation
- show the RIP table for nodes

Figure1 shows an overview of our e-Learning system for network study. This is Japanese User Interface version. At the right-hand sub window shown in Figure1, the system automatically generates an according table for Routing Information Protocol(RIP table) to define IP routing rules, namely IP-based packet transfer between IP routers and sources/destinations which have been specified by users (learners). Learner, especially beginners of network-related subjects, can recognize IP routing detail from the system in the above way. In other words, learners can understand behavior of IP routing without detail specification of the relevant IP routing rules.

2.2 Modified JavaScript Code for Multilingual Facilities

We are asked to improve an existing e-Learning system for network study into a new one with Multilingual Supports in order to

Figure 2. JavaScript source for MultiLingual Supports.

be available not only for Japanese but also for French/Portuguese spoken users as well as English spoken one. After investigation of the relevant system, we have decided to employ UTF-8 as description code for HTML and to modify the adequate expression in the JavaScript source for our new improvement. Figure2 shows a part of the modified JavaScript source for the newly modified e-Learning system with MultiLingual facilities for foreign users to learn network design and demonstration.

After implementation of MultiLingual Supports on JavaScript source, a newly modified e-Learning system can provide a menu on the middle of UI to select Language types for suitable learning among Japanese, English, French and Portuguese. Figure3 shows such a menu on the middle of UI of the system. With our multilingual supports, users of this system can select their favorite language types for their effective and efficient learning.

3 MULTILINGUAL IMPROVEMENT

This section illustrates modified user interface for English, French and Portuguese spoken learners at first, and then demonstrates multi-

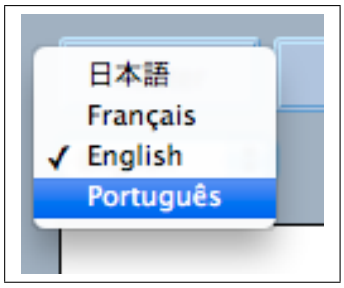


Figure 3. Language Selection for Suitable Learning.

lingual mode to design network and to animate (=simulate) IP routing behavior secondarily.

3.1 Addition of French/English Interface

We have already prepared user interface for English/French/Portuguese users as well as Japanese ones in order for world-wide learners to understand our newly modified system quickly and manipulate it smoothly in a relatively shorter period than used to be. Figure4 shows our new user interfaces for English spoken users (a:upper) and for French spoken ones (b:lower, “Français”). We have decided to

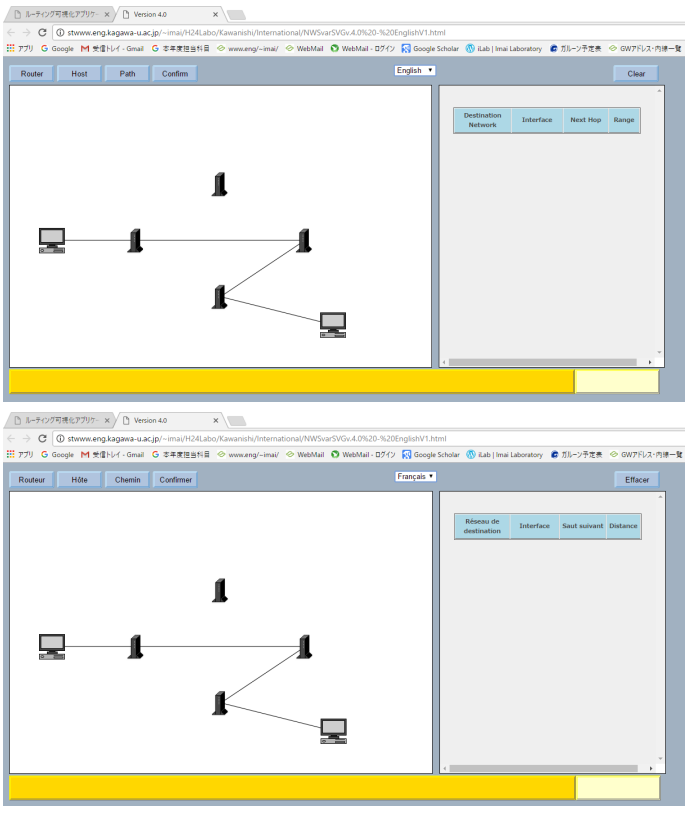


Figure 4. User Interface for English(a:upper)/ French(b:lower) Learners.

ISBN: 978-1-941968-42-0 ©2017 SDIWC

specify user interface of our system for English users as a default due to world-level standard. For Japanese, English, French and/or Portuguese spoken users, our newly improved system will provide almost the same facilities and services for its learners to understand network structure and behavior through network topology design and animation of IP-routing. All of us would be very happy if East-South Asian young students, Central African young students and/or Brazilian young students were interesting in our newly modified system and were willing to learn network mechanism, protocols and future applications through their handling our system.

3.2 MultiLingual Explanation of Usage based on IntroJS

We have employed IntroJS to provide smart MultiLingual Explanation of Usage in order for users of the world to recognize how to manipulate our system efficiently. Figure5 shows an example of multilingual explanation of usage in Portuguese.

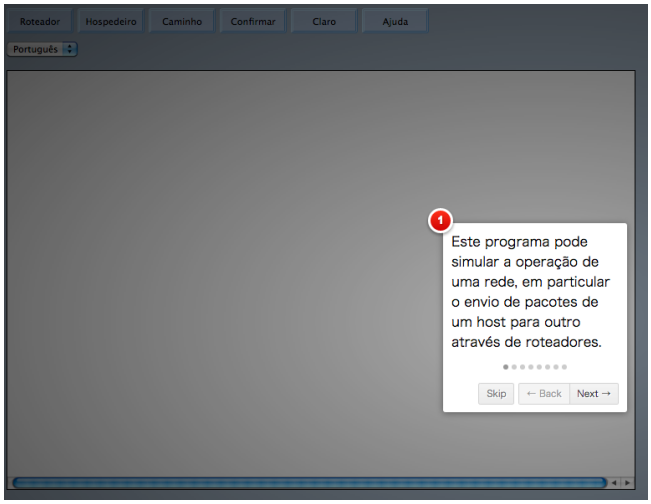


Figure 5. IntroJS-based MultiLingual Explanation of Usage in Portuguese.

Users can select types of language in the way shown in Figure3, and then our system automatically provides the corresponding type of multilingual explanation for them, for example, such as IntroJS-based usage, guideline and prompt, the RIP table, annotation generated with mouse-over manipulation and so on.

3.3 Multilingual Mode of Manipulation (Design/Animation)

Even at the mode of French user interface, after finishing designing of network topology as is shown in Figure4, a user is asked to point out source node (=terminal at the main window(the left hand) of UI in Figure6. He/She specifies the leftmost node as source by means of mouse. Source is specified by a red rounding circle shown in Figure6.

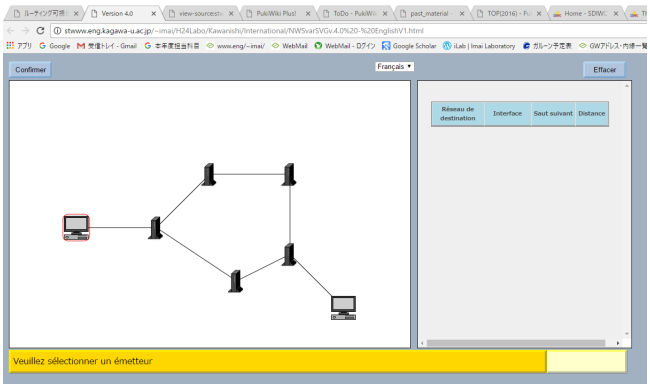


Figure 6. Design and Animation of Network for French Learners No1.

After he/she points out destination node by mouse, the system also generates a red circle around the specified node as confirming signal shown in Figure7. Now he/she selects favorite speed type among three for animation control of IP routing shown in Figure8 in order to recognize how an IP packet flows from source node to destination one through the shortest path of network connectivity.
Figure9 shows a screen shot at the second

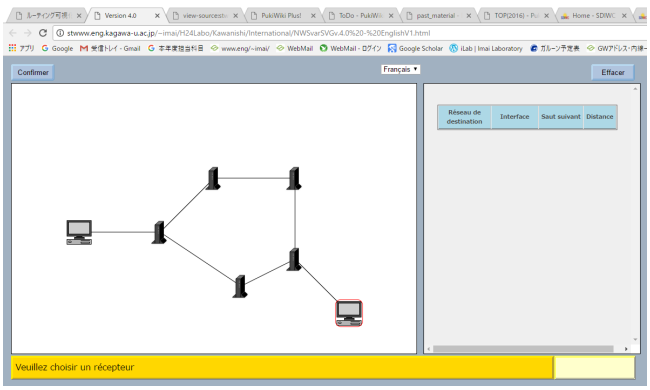


Figure 7. Design and Animation of Network for French Learners No2.

ISBN: 978-1-941968-42-0 ©2017 SDIWC

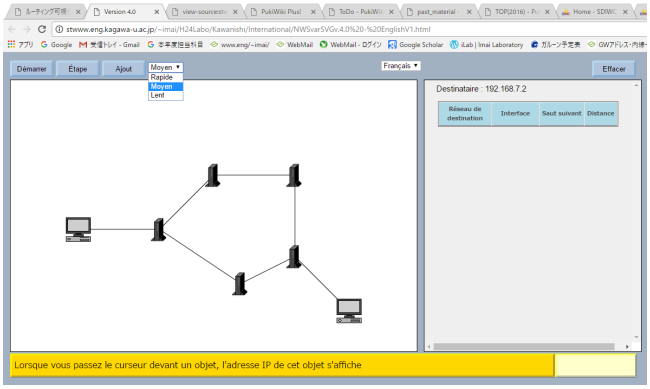


Figure 8. Design and Animation of Network for French Learners No3.

router receiving the packet from the first one and then transferring it to the third one. The relevant second router has more than two different IP addresses for realization of flexible connectivity, whose values of addresses are shown in the red numbers and the router itself is specified by the red rounding circle. And he/she can find the related table for Routing Information Protocol in the right hand window of the user interface. He/She can also check the value of TTL(Time to live) in the bottom yellow line of the same interface.

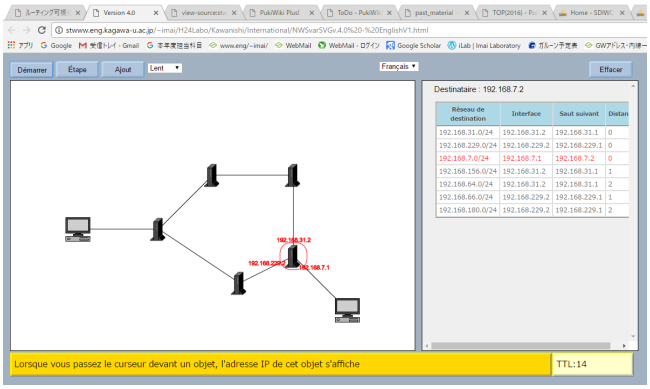


Figure 9. Design and Animation of Network for French Learners No4.

Not only French spoken learners but also English/Portuguese/Japanese spoken learners can suitably manipulate our system according to the French/English/Portuguese/Japanese guidance(s) so that they will be able to recognize how an IP packet flows from source to destination and how each router receives and transfers the relevant packet in a coordinated way which is more smoothly recognized by learners through reference of the table for Routing Information Protocol.

4 ADVANTAGE IN REALIZATION OF MULTILINGUAL SYSTEM

This section discusses advantage in realization of multilingual e-Learning system based on international internship program.

4.1 Improvement of the Existing System and Application

People say that Kagawa university has been wanting to resolve its several problems, for example, its faculty members are willing to enhance their research themes and/or apply their research results into larger/other domains than used to be. As they consider this time international internship program to be a good trigger, they have requested us to improve their existing systems into world-wide available/applicable one together with us. And they seem to be pleased after we can understand their existing system in a short period and propose our idea to improve their system by means of multilingual services.

For us, it is an attractive chance and useful opportunity to have good experience to investigate existing systems by Japanese Master-course students, namely check the knowledge and performance levels of such students. And moreover we feel lucky because we can choose our favorite target(s) among their preparing tasks and/or projects. This is why we had wanted to build Web-based application, especially program written in JavaScript as international project of internship program before our visiting to Japan.

4.2 Collaboration through Multilingual Viewpoints

Kagawa university has assigned some master-course students to be supporters and tutors for applicants of international internship program and asked them to help us to be living and studying alone during staying. They are

ISBN: 978-1-941968-42-0

good friends but seem to be not good at speaking French as well as English, while we were learning Japanese in our country so that we are not so difficult to communicate them directly with their foreign languages, namely English, French or Portuguese. In such a case, collaboration among internship foreign students and Japanese ones used to be not so easy to accomplish.

This time we are happy because not only we but also they can understand and manipulate "GitLab/GitHub" so that it is good for us and them to accomplish fruitful collaboration through Git-based environment. They have provided their existing system to us by means of their GitLab, and we obtain from GitLab its detail of such system(s) at the source code level and improve it(them) into world-wide available one(s) together with them during internship program.

And we have another good experience to introduce and explain our national education scheme from elementary school to university level in English among Japanese students and teachers. Not only they but also we have obtained good stimulation from such introduction an discussion in English. We hope that we will have additional research themes through those discussion among all the participants.

5 CONCLUSION

This paper describes detail about our multilingual improvement of existing e-Learning system of Kagawa university. The existing e-Learning system was prepared for learners to understand network connectivity and behavior of IP-packet routing by means of its visualization facility. But that system had been available for Japanese spoken user only because of explanation written in Japanese and not special guidance for foreign users.

With our multilingual supports, not only Japanese spoken learners but also English/French/Portuguese spoken learners have been able to manipulate improved e-Learning system more effectively and smoothly than used to be. And moreover they will be able to recognize how to design sample network topology and how to perform demonstration of IP-packet routing from specified source to specified destination by means of animation by means of section of user interface which can provide English/French/Portuguese supports for learners.

Our team is beginning to write the contents of questionnaire for users to answer after manipulation of the system. They are described not only in Japanese but also in French/English/Portuguese in order to obtain certain evaluation from users of the world. We are ready to carry out questionnaire for French/English/Portuguese spoken users since it was done for Japanese.

ACKNOWLEDGEMENTS

We are very thankful to International Internship Program between ESIEE Paris and Kagawa university, Japan because of its providing very useful opportunity for us to staying in Japan and accomplish a creative task.

REFERENCES

[1] M. Arai, et al., "Development and Evaluation of TCP/IP Protocol Learning Tools (in Japanese)," IPSJ Journal, vol.44, no.12, pp.3242–3251, Dec. 2003.

[2] Y. Tateiwa, et al.,"Development of a system to visualize computer network behavior for learning to associate LAN construction skills with TCP/IP theory, based on virtual environment software (in Japanese)," IPSJ Journal, vol.48, no.4, pp1684–1694, April, 2007.

[3] C. Kawanishi, et al., "Development and Evaluation of Learner-centric Graphical Educational Tool for Network Study," Proc. of 2013 International Conference on Humanized Systems (ICHS2013@Takamatsu), pp.108–113, Sept. 2013.

[4] Y.Imai, C.Kawanishi, T.Hattori, Y.Hori, "Development and Evaluation of Learner-centric Graphical Educational Tool for IP Routing and Network Behavior," IEEJ(the Institute of Electrical Engineers of Japan) Transactions on Electronics, Information and Systems, Vol.134, No.11, pp.1634–1639, Nov. 2014.

A Study for effectiveness of Dimensionality Reduction for State-action Pair Prediction –Training set reduction using Tendency–

Masashi Sugimoto[1*] Naoya Iwamoto[1] Robert W. Johnston[1]
Keizo Kanazawa[1] Yukinori Misaki[1] Kentarou Kurashige[2]

[1] National Institute of Technology, Kagawa College

[2] Muroran Institute of Technology

*Department of Electronic Systems Engineering,
551 Kohda, Takuma, Mitoyo, Kagawa, 769-1192 Japan.

Email(corresponding author): sugimoto-m@es.kagawa-nct.ac.jp

ABSTRACT

This paper investigates the effectiveness of reduction of training sets and kernel space for action-decision using future prediction. Considering a working in a real environment based on future prediction, it's necessary to know the property of its state and disturbance that will be given by the outside environment. On the other hand, obtaining the property of the disturbance depends on specification for target processor, especially, sensor resolution or processing ability of the processor. Therefore, sampling rate settings will be limited by hardware specification. In contrast, in case of a future prediction using a machine learning, it predicts that based on the tendency that obtained by past training or learning. In this kind of situation, the learning time will be proportionally larger to training data. At worst, the prediction algorithm will be hard to work in real time due to time-complexity.

In the proposed method, the possibility of carefully analyzing the algorithm and applying dimensionality reduction techniques in order to accelerate the algorithm has been considered. In particular, we will consider that to reduce the training sets and kernel space based on the recent tendency of disturbance or state using FFT and pattern matching will be focused on. From this standpoint, we will propose the method that to dimensionality reduction dynamically based on the tendency of disturbance.

KEYWORDS

Online SVR, Predict and Control using State-action Pair Prediction, Dimensionality Reduction

1 INTRODUCTION

Considering an action decision based on future prediction, it's necessary to know the property of disturbance that will be given by outside environment [1]. On the other hand, obtaining the property of the disturbance is depend on specification for target processor, especially, sensor resolution or processing ability of the processor. Therefore, sampling rate settings will be limited by hardware specification. In contrast, in case of a future prediction using a machine learning, it predicts that based on the tendency that obtained by past training or learning. In this kind of situation, the learning time will be proportionally larger to training data [1, 2].

A State-action Pair Prediction had been proposed. In this method, the prediction performance [3] and action decision methods [4, 5], had been considered based on some prediction results. In above-mentioned methods, the behavior of the robot has been considered when an unknown periodic disturbance signal will be given the robot, continuously. On the other hand, in these studies, the learning space had not been considered in action decision or future prediction. In general, non-linear clustering (or regression, such as this work), Kernel Function was used, that allows growth of the SVM (also SVR) solution, which starts invading other space, and this "other space" is called the Features Space. This allows us to change the information from one linear space

to another one. This permits us to better classify (or regression) the examples. However, the speed of learning depends mostly on the number of support vectors, that can influence significatively performances. Therefore, in simply, the complexity of learning will be proportional as the size of training sets. If the recent tendency of disturbance or state, or these period will be obtained, the training sets will be reduced. Moreover, the length of training sets will be fixed in spite of a new training set will be added.

Therefore, in the proposed method, the possibility of carefully analyzing the algorithm and applying dimensionality reduction techniques in order to accelerate the algorithm has been considered. In particular, we will consider that to reduce the training sets and kernel space based on the recent tendency of disturbance or state using FFT and pattern matching will be focused on. From this standpoint, we will propose the method that to dimensionality reduction dynamically based on the tendency of disturbance.

This paper is organized as follows: In section II, how to reduce a learning space (feature space) dynamically, will be motivated. Further, details about the decide a learning space based on the Nearest-neighbor one-step-ahead forecasts and Nyquist-Shannon Sampling Theorem will be stated. In Section III, a verification experiment configuration will be described. In Section IV, the summary of this work is concluded.

2 AN APPROACH OF THE REDUCING THE LEARNING SPACE BASED ON FREQUENCY PROPERTY OF THE DISTURBANCE SIGNAL

2.1 About Former Our Works

We mentioned in citations that for controlling the robot in a dynamic environment, it can realize choosing the action that adopted the current result by predicting the future state using previous actions and states. In this paper, we will try to consider that obtain the optimal action that is minimizing the body pitch angle of the inverted pendulum, in case of continuing the predictive disturbance using the prediction the State-action Pair that had proposed in the former our study [3]. Therefore, in this paper, we considering the system that decides the action to optimize as the proposed method in fig. 1 based on the former study and the study [4].

In fig. 1, this system will be applied optimal

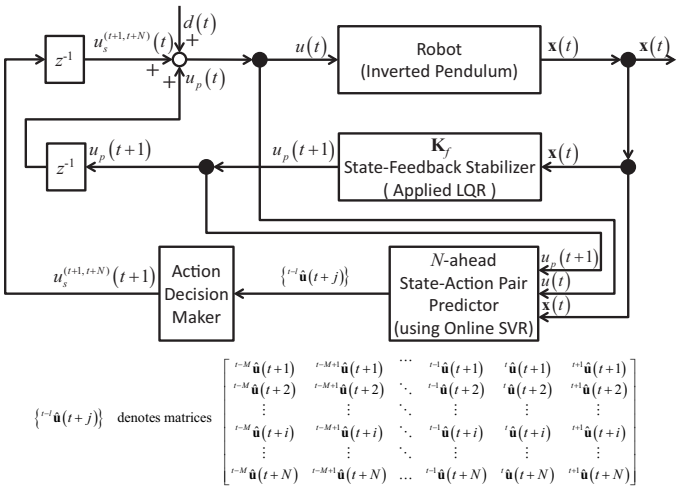


Figure 1. A Outline of the Deciding the Optimal Action for the Robot using the Prediction of State-action Pair [5]

control using a gain $\mathbf{K}_f$ as an optimal feedback gain, and in parallel, deciding the action that will have to take in the future using the prediction of state-action pair. In fig. 1, ${}^{t-l}\hat{\mathbf{u}}(t+j)$ is describing the prediction result of the control input $\mathbf{u}(t+j)$, when that input predicted in time $(t-l)$. And hence, this proposed method is revised the current action using the action that combining the optimal control and the prediction result of State-action Pair Prediction. The structure of the prediction of State-action Pair is named "N-ahead State-action Pair Predictor," that the internal structure is described in fig. 2 [3]. The proposed method obtain a se-

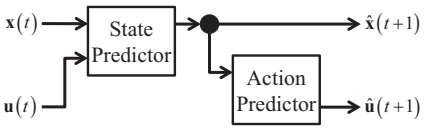


Figure 2. Outline of the Prediction System of State and Action [3]

ries of action in time $(t+N)$ in the distant future from current time t using N-ahead State-

action Pair Predictor. Now from this prediction series, the current action, combining "*the action will be taken in the future*" and using the prediction series of the action, will be able to revise.

Hereby, in this system, the optimal compensation control input $\mathbf{u}(t)$ will be given as follow:

$$\mathbf{u}(t) = \mathbf{u}_p(t) + \mathbf{u}_s^{(t+1,t+N)}(t) + \mathbf{d}(t) \qquad (1)$$

In this equation, $\mathbf{u}_p(t)$ denotes an optimal control action with optimal feedback control gain, $\mathbf{u}_s^{(t+1,t+N)}(t)$ is generated from "Action Decision Maker," and $\mathbf{d}(t)$ denotes unknown periodic disturbance input signal. Then, $\mathbf{u}_s^{(t+1,t+N)}(t)$ can be defined as follows.

$$\mathbf{u}_s^{(t+1,t+N)}(t) = \sum_{i=1}^{N} \alpha_i \hat{\mathbf{u}}(t+i) \qquad (2)$$

Moreover, in this study, the coefficient α_i will be defined as:

$$\alpha_i = \frac{N+i-1}{100 \cdot N} \qquad (3)$$

From this technique, we create an action that can correspond in ahead a time, can obtain the optimal action that will be considered in the future.

2.2 Basic Idea

In Section I, we stated about the relationship dimensionality reduction and size of training sets for the learning and predicting. From this viewpoint, we can reduce the training time and the size of the learning space for predicting future states and action if we can reduce the training sets.

In SVR, it defines "support vector," and a number of them are less than numbers of other training sets. In detail, support vector denotes the property of unknown function. From this viewpoint, it's not needed a precision training set for unknown function. Thus, removing any training samples that are not relevant to support vectors might have no effect on building the proper decision function [6]. In other words, in this study, the training sets can reduce if the unknown periodic disturbance will be applied to plant model. In addition, the model can be built a prediction model for an almost-periodic disturbance signal if the support vectors can denote the property of an almost-periodic; that is, in the proposed method, the training sets will be reduced based on the recent tendency of disturbance or state, or these period. Moreover, the length of training sets will be fixed in spite of a new training set will be added. Moreover, the support vectors of "one-period" of almost-periodic disturbance will be used repeatedly. Therefore, in the proposed method, the prediction model can be predicted and adapted the plant if an unknown periodic disturbance will be applied as same as former works, in spite of dimensionality reduction dynamically based on the tendency of disturbance.

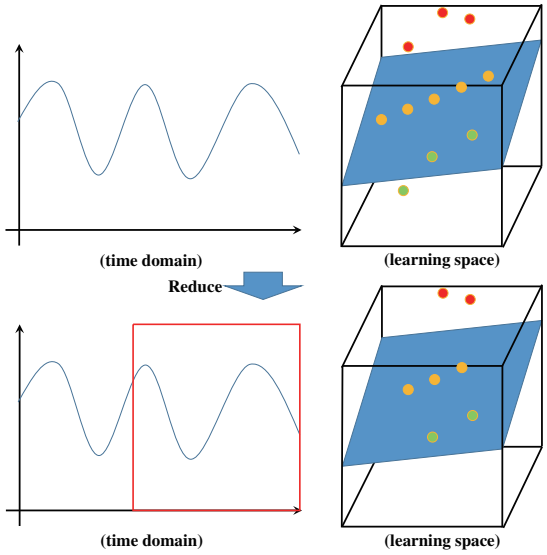


Figure 3. Outline of reduce the learning space according to a Period of Disturbance Signal

2.3 How to Estimate the Frequency of A Disturbance Signal and Reduce Learning Space

As mentioned above, the unknown periodic signal will be used as a disturbance signal. Therefore, we will try to analyze the property of disturbance signal, to reduce the learning space (in fig. 3). In this case, the disturbance signal will be represented as a similar tendency as a pattern.

From this property, the Nearest-neighbor One-step-ahead forecasts [7] will be applied, to detect a cycle period of the disturbance signal. Let illustrate about the Nearest-neighbor One-step-ahead forecasts. As shown in fig. 4, target function $f(t)$ will be repeated similar tendency. In this case, we want to predict at time $t+1$ the next value of the series f. The pattern $f(t-12), f(t-6)$ is the most similar to the pattern. Then, the prediction will be calculated. As a result, the Nearest-neighbor One-

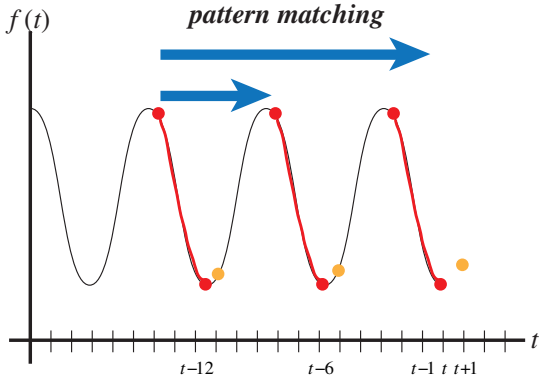


Figure 4. Outline of Nearest-neighbor One-step-ahead Forecasts

step-ahead forecasts provide not only a one-step prediction result, but also the cycle period. From these results, the cycle period $T_{\text{disturbance}}$ will be calculated as follows:

$$T_{\text{disturbance}} = |t_{\text{max, disturbance}} - t_{\text{min, disturbance}}| \times 2 \tag{4}$$

Here, $t_{\text{max, disturbance}}$ denotes the time when the maximum values of disturbance signal $d(t)$ has been reached, moreover, $t_{\text{min, disturbance}}$ denotes the time when the minimum values of disturbance signal $d(t)$ has been reached.

On the other hand, we have stated that we will try to analyze the property of disturbance signal, to reduce the training sets, as mentioned above. Therefore, Nyquist-Shannon Sampling Theorem will be focused on, to reduce the training sets, and to keep the property of the disturbance signal. Now, the Theorem states: *A sufficient sample-rate is therefore $2B$ samples/second, or anything larger. Equivalently, for a given sample rate f_s, perfect reconstruction is guaranteed possible for a bandlimit* $B < f_s/2$. From this theorem, the sampling rate t'_s will be defined as follows:

$$t'_s \leq \frac{T_{\text{disturbance}}}{2} \tag{5}$$

In here, the sampling rate t'_s ts will be integral multiple of original sampling rate t_s. Then, a result obtained by divide t'_s by t_s will be training set that build a prediction model. Therefore, new training sets will be defined:

$$N = \frac{t'_s}{t_s} \tag{6}$$

$$S = \{s_{t-N}, s_{t-N+1}, \ldots, s_t\} \tag{7}$$

In above equation, S is a list of support sets.

The proposed method predicts events in that is given in the training sets, however, does not reduce former training sets. In this section, how to implement the future prediction will be stated.

In this case, a next state $\hat{x}_{t+1,i}, i \in \dim \hat{\mathbf{x}}_{t+1}$ (i denotes an element of all the robot's state) is estimated by using the state and action are defined by $\mathbf{z}_t = \begin{bmatrix} x_{t,1} & \ldots & x_{t,n} & | & a_t \end{bmatrix}$. Therefore, this vector $\mathbf{z}_t$ is an $(n+1)\times 1$ vector. Next, let's consider the sum-of-squares error function J_S from training set $\{\mathbf{x}_j, y_j\}$ described by the SVR model $y(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) + b$ [8].

$$J_S(\mathbf{w}) = \frac{1}{2} \sum_{j=t-N}^{t} \left\{ \mathbf{w}^\top \phi(\mathbf{x}_j) + b - y_j \right\}^2 + \frac{\lambda}{2} \mathbf{w}^\top \mathbf{w} \quad (\lambda \geq 0) \tag{8}$$

where $\mathbf{w}^\top$ indicates the transpose of $\mathbf{w}$. Here, λ represents the regularization parameter, and $\mathbf{w}$ represents the weight matrix of the SVR model. The weight matrix $\mathbf{w}$ is found by setting the gradient for minimizing the sum-of-squares error function J_S to zero (thus,

$\partial J_S(\mathbf{w})/\partial \mathbf{w} = 0$). Hence,

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}} J_S(\mathbf{w}) =& 2 \times \frac{1}{2} \sum_{j=t-N}^{t} \left[\left\{ \mathbf{w}^\top \phi(\mathbf{x}_j) + b - y_j \right\} \right. \\ & \left. \phi(\mathbf{x}_j) \right] + \frac{\lambda}{2}\mathbf{w} + \frac{\lambda}{2}\mathbf{w} = 0 \\ 0 =& \sum_{j=t-N}^{t} \left[\left\{ \mathbf{w}^\top \phi(\mathbf{x}_j) + b - y_j \right\} \right. \\ & \left. \phi(\mathbf{x}_j) \right] + \lambda \mathbf{w} \\ \mathbf{w} =& -\frac{1}{\lambda} \sum_{j=t-N}^{t} \left\{ \mathbf{w}^\top \phi(\mathbf{x}_j) + b - y_j \right\} \phi(\mathbf{x}_j) \\ =& \sum_{j=t-N}^{t} a_j \phi(\mathbf{x}_j) = \Phi^\top \mathbf{a} \end{aligned} \tag{9}$$

$$\textit{where } \mathbf{a} = \begin{bmatrix} a_{t-N} & \dots & a_t \end{bmatrix}^\top, \quad a_j = -\frac{1}{\lambda} \left\{ \mathbf{w}^\top \phi(\mathbf{x}_j) + b - y_j \right\}$$

Now, Φ is called the design matrix, and the j-th row is described by $\phi(\mathbf{x}_j)^\top$. Here, the parameter vector $\Phi\mathbf{a}$ replaces $\mathbf{w}$,

$$\begin{aligned} J(\mathbf{a}) =& \frac{1}{2}\mathbf{a}^\top \Phi\Phi^\top \Phi\Phi^\top \mathbf{a} - \mathbf{a}^\top \Phi\Phi^\top \mathbf{y} \\ & + \frac{1}{2}\mathbf{y}^\top \mathbf{y} + \frac{\lambda}{2}\mathbf{a}^\top \Phi\Phi^\top \mathbf{a} \end{aligned} \tag{10}$$

Now, the Gramian matrix $\mathbf{K} = \Phi\Phi^\top$ will be defined. Here, the matrix coefficient of $\mathbf{K}$ is given by

$$K_{jm} = \phi(\mathbf{x}_j)^\top \phi(\mathbf{x}_m) = k(\mathbf{x}_j, \mathbf{x}_m) = Q_{jm} \tag{11}$$

This matrix coefficient is the symmetric matrix as a kernel matrix. Now, let's rearrange the sum-of-squares error function J_S by using the Gramian matrix:

$$\begin{aligned} J_S(\mathbf{a}) =& \frac{1}{2}\mathbf{a}^\top \mathbf{K}\mathbf{K}\mathbf{a} - \mathbf{a}^\top \mathbf{K}\mathbf{y} \\ & + \frac{1}{2}\mathbf{y}^\top \mathbf{y} + \frac{\lambda}{2}\mathbf{a}^\top \mathbf{K}\mathbf{a} \end{aligned} \tag{12}$$

The equation is rearranged by isolating $\mathbf{a}$:

$$\mathbf{a} = (\mathbf{K} + \lambda \mathbf{I}_N)^{-1} \mathbf{y} \tag{13}$$

Here, $\mathbf{I}_N$ represents the $N \times N$ identity matrix. Therefore, the prediction result $\hat{y}(\mathbf{x})$ for the SVR model to input $\mathbf{x}$ can be derived the equation anew as

$$\begin{aligned} \hat{y}(\mathbf{x}) &= \mathbf{w}\phi(\mathbf{x}) + b = \mathbf{a}^\top \Phi\phi(\mathbf{x}) + b \\ &= \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \lambda \mathbf{I}_N)^{-1} \mathbf{y} + b \end{aligned} \tag{14}$$

$$\text{where} \quad \mathbf{k}(\mathbf{x}) = \begin{bmatrix} k(\mathbf{x}_{t-N}, \mathbf{x}) & \dots & k(\mathbf{x}_j, \mathbf{x}) \end{bmatrix}^\top$$

In this time, prediction result and Kernel matrix will be updated as below:

$$\begin{aligned} \hat{y}(\mathbf{x}) &= \mathbf{w}\phi(\mathbf{x}) + b = \mathbf{a}^\top \Phi\phi(\mathbf{x}) + b \\ &= \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \lambda \mathbf{I}_N)^{-1} \mathbf{y} + b \end{aligned} \tag{15}$$

$$\text{where} \quad \mathbf{k}(\mathbf{x}) = \begin{bmatrix} k(\mathbf{x}_{t-N}, \mathbf{x}) & \dots & k(\mathbf{x}_t, \mathbf{x}) \end{bmatrix}^\top$$

$$Q = \begin{bmatrix} Q_{s_{t-N}, s_{t-N}} & \cdots & Q_{s_{t-N}, s_t} \\ \vdots & \ddots & \vdots \\ Q_{s_t, s_{t-N}} & \cdots & Q_{s_t, s_t} \end{bmatrix} \tag{16}$$

The matrix Q contains the values of kernel function and it is called kernel matrix. In these equations, learning space Q and training sets $\mathbf{x}$ will be re-construct each sampling time and adding new training data. Therefore, the learning space and training sets will be reduced each sampling time. As mentioned before, the speed of learning depends mostly on the number of support vectors, that can influence significatively performances. As a result, the speed of learning will be improved than former works.

3 THE VERIFICATION EXPERIMENT – COMPUTATIONAL SIMULATION USING THE PROPOSED METHOD

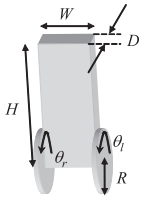
3.1 Outline of the Experiment

In this experiment, we stabilize the posture of a two-wheeled self-propelled inverted pendulum "NXTway-GS" (fig. 5) as an application, using the computer simulation. In this verification experiment, we compared the control response of the proposed method with the ordinary method. Furthermore, in proposed method, the predictor only used the proximate predicted result repeatedly training data from 0 [s] (don't reduce) or training sets that reduced in each sampling time, for postural control.

3.2 Simulation Setup - the NXTway-GS Model

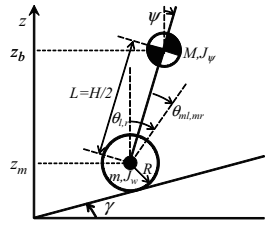
NXTway-GS (fig. 5) can be considered as inverted pendulum model shown in fig. 6. Figure 6 shows the side view and the plane view of the model. The coordinate system used in 3.3 is described in fig. 6. In figure 6, ψ denotes the body pitch angle and $\theta_{ml,mr}$ denotes the DC motor angle (l and r indicate *left* and *right*). The physical parameters of NXTway-GS are listed in table 1.

(a) NXTway-GS

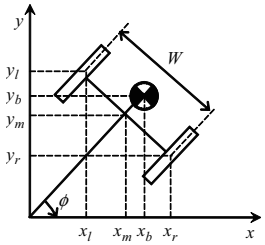


(b) Model of NXTway-GS

Figure 5. Two-wheeled Inverted Pendulum "NXTway-GS"



(a) Side view



(b) Plane view

Figure 6. The Side View and the Plane View of NXTway-GS [10]-[11]

3.3 Simulation Setup - Modeling of the NXTway-GS

We can derive the equations of motion of the inverted pendulum model using the Lagrange equation based on the coordinate system in fig. 6. If the direction of model is in the x-axis positive direction at $t = 0$, the equations motion for each coordinate are given as ([10]-[11]) ;

$$\left[(2m+M)R^2 + 2J_w + 2n^2J_m\right]\ddot{\theta} + (MLR - 2n^2J_m)\ddot{\psi} - Rg(M+2m)\sin\gamma = F_\theta \tag{17}$$

$$(MLR - 2n^2J_m)\ddot{\theta} + \left(ML^2 + J_\psi + 2n^2J_m\right)\ddot{\psi} - MgL\psi = F_\psi \tag{18}$$

$$\left[\frac{1}{2}mW^2 + J_\phi + \frac{W^2}{2R^2}\left(J_w + n^2J_m\right)\right]\ddot{\phi} = F_\phi \tag{19}$$

Here, we consider the following variables $\mathbf{x}_1, \mathbf{x}_2$ as the state variables and $\mathbf{u}$ as the input variable ($\mathbf{x}^\top$ indicates the transpose of $\mathbf{x}$.).

$$\mathbf{x}_1 = \begin{bmatrix}\theta & \psi & \dot{\theta} & \dot{\psi}\end{bmatrix}^\top \tag{20}$$

$$\mathbf{x}_2 = \begin{bmatrix}\phi & \dot{\phi}\end{bmatrix}^\top \tag{21}$$

$$\mathbf{u} = \begin{bmatrix}v_l & v_r\end{bmatrix}^\top \tag{22}$$

Consequently, we can derive the state equations of the inverted pendulum model from eq. (17), (18) and (19).

$$\frac{d}{dt}\mathbf{x}_1 = \mathbf{A}_1\mathbf{x}_1 + \mathbf{B}_1\mathbf{u} + \mathbf{S} \tag{23}$$

$$\frac{d}{dt}\mathbf{x}_2 = \mathbf{A}_2\mathbf{x}_2 + \mathbf{B}_2\mathbf{u} \tag{24}$$

In this paper, we only use the state variables $\mathbf{x}_1$. Because $\mathbf{x}_1$ is including body pitch angle as important variables ψ and $\dot{\psi}$ for control of self-balancing, and we will not consider plane motion ($\gamma_0 = 0, \mathbf{S} = \mathbf{0}$).

3.4 Simulation Setup - How to Apply the Online SVR to the State Predictor

In this method, we use Online SVR [9] as a learner. Moreover, we applied RBF kernel [13] as the kernel function to the Online SVR of the learner. The RBF kernel on two samples $\mathbf{x}$ and $\mathbf{x}'$, represented as feature vectors in some input space, is defined as

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\beta\left|\left|\mathbf{x} - \mathbf{x}'\right|\right|^2\right) \tag{25}$$

And the learning parameters of Online SVR are listed in table 2. In table 2, $i \in \{1, 2, 3, 4\}$.

Table 1. Physical Parameters of the NXTway-GS

Symbol	Value	Unit	Physical property
g	9.81	[m/s^2]	Gravity acceleration
m	0.03	[kg]	Wheel weight [10]
R	0.04	[m]	Wheel radius
J_w	$\frac{mR^2}{2}$	[kgm^2]	Wheel inertia moment
M	0.635	[kg]	Body weight [10]
W	0.14	[m]	Body width
D	0.04	[m]	Body depth
H	0.144	[m]	Body height
L	$\frac{H}{2}$	[m]	Distance of center of mass from wheel axle
J_ψ	$\frac{ML^2}{3}$	[kgm^2]	Body pitch inertia moment
J_ϕ	$\frac{M(W^2+D^2)}{12}$	[kgm^2]	Body yaw inertia moment
J_m	1×10^{-5}	[kgm^2]	DC motor inertia moment [11]
R_m	6.69	[Ω]	DC motor resistance [12]
K_b	0.468	[V·s/rad.]	DC motor back EMF constant [12]
K_t	0.317	[N·m/A]	DC motor torque constant [12]
n	1	[1]	Gear ratio [11]
f_m	0.0022	[1]	Friction coefficient between body and DC motor [11]
f_W	0	[1]	Friction coefficient between wheel and floor [11]

Table 2. Learning Parameters of the Online SVR

Symbol	Value	Property
C_i	300	Regularization parameter or predictor of x_i
ϵ_i	0.02	Error tolerance for predictor of x_i
β_i	30	Kernel parameter for predictor of x_i

3.5 Simulation Setup - How to Apply the Linear-quadratic Regulator to the Action Predictor

In this experiment, we apply LQR (Linear-quadratic Regulator) as an action prediction (And a predictor). So we design the controller as an action predictor based on modern control theory. This LQR calculates the feedback gain $\mathbf{k}_f$ so as to minimize the cost function J_C given as the following;

$$J_C = \int_0^\infty \left[\mathbf{x}^\top(t)\mathbf{Q}\mathbf{x}(t) + \mathbf{u}^\top(t)\mathbf{R}\mathbf{u}(t)\right] dt \tag{26}$$

The tuning parameter is the weight matrix for state $\mathbf{Q}$ and for input $\mathbf{R}$. In this paper, we choose the following weight matrix $\mathbf{Q}$ and $\mathbf{R}$;

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 6 \times 10^5 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 4 \times 10^2 \end{bmatrix} \tag{27}$$

$$\mathbf{R} = 1 \times 10^3 \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{28}$$

Then, we obtain the feedback gain $\mathbf{k}_f$ from minimizing J_C. Therefore, we apply $\mathbf{k}_f$ as an action predictor [3]. And hence, in this experiment, we do not consider the plane move of the two-wheeled inverted pendulum. In other words, we consider that $\phi = 0, \theta_{ml} = \theta_{mr}$, and $\mathbf{u} = u, \mathbf{d}(t) = d(t)$.

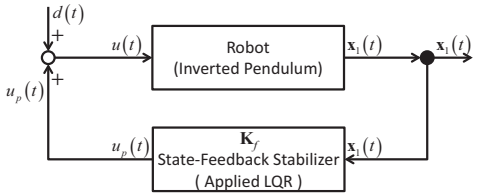


Figure 7. Control Input Obtained by Mixing the Action and Disturbance Inputs

3.6 Conditions of Simulation - Acquiring the Training Sets

In this experiment, we mix the action signal with a known disturbance signal $\mathbf{d}(t)$ (figs. 7, 8, and 9), and $\mathbf{d}(t)$ is given as

$$d(t) = A_{d1} \sin(2\pi f_{d1} t) \tag{29}$$

Then, the signal $\mathbf{d}(t)$ will be mixed to the

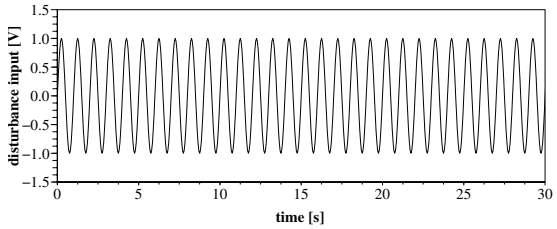


Figure 8. Disturbance Signal in Control Inputs $\mathbf{d}(t)$ (overall)

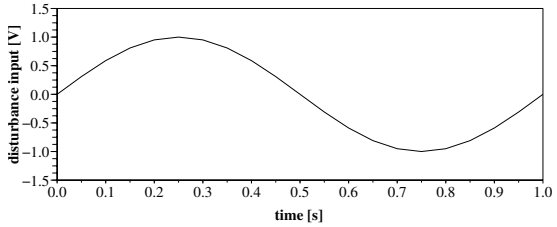


Figure 9. Disturbance Signal in Control Inputs $\mathbf{d}(t)$ (focused on 0 [s] to 1[s])

model. Herewith, we can acquire the training sets from the two-wheeled inverted pendulum. In figures 10 to 12 shows training sets that were obtained from the computer simulation of the stabilize control of the two-wheel inverted pendulum. Moreover, the properties of disturbance that we provide as input and other conditions of a simulation are listed in table 3.

3.7 Simulation Results

Figures 10 and 11 show compensation results of the state of $\mathbf{x}_1$, and fig. 12 shows the prediction of the control input and compensation input using prediction result of $\mathbf{u}$.

In this section, we will not consider the part that is given in real training sets. Thus we will only argue and focus on the part of the graph pertaining to the state predicted part shown in T (at $t = 3.00$ [s]) of figs. 10 and 11.

3.8 Discussion on Simulated Results

In here, starting and predicting the state predicted point is shown at $t = 3.00$ [s] as shown in T.

According to these results (figs. 10 through 12), compensation results using the proposed

Table 3. Parameters for Condition of a Simulation

Sybmol	Value	Unit	Physical property
ψ_0	0.0262	[rad.]	Initial value of body pitch angle
γ_0	0.0	[rad.]	Slope angle of movement direction
t_s	0.05	[s]	Sampling rate
$t_{d,\text{start}}$	0.0	[s]	Start time of application of predictable disturbance
$t_{d,\text{finish}}$	45.0	[s]	Finish time of application of predictable disturbance
A_{d1}	1.0	[V]	Amplitude of predictable disturbance
f_{d1}	1.0	[Hz]	Frequency of predictable disturbance
N_s	60	—	Initial dataset length
$N_{\max}$	241	—	Maximum dataset length for the prediction
N	20	—	Step size of outputs for N-ahead State-action Pair Predictor's outputs
α_i $i \in N$	$\frac{N-i+1}{100N}$	—	Weight coefficients for $\hat{u}(t+i), i \in N$ (for the contrast experiment [4])

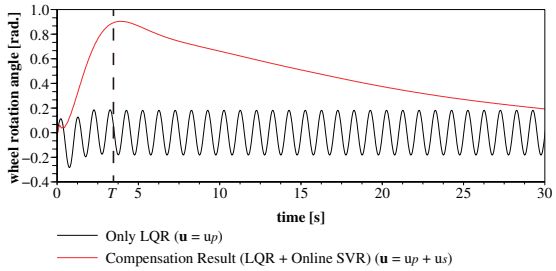


Figure 10. Control Response of the Wheel Rotation Angle θ (1)

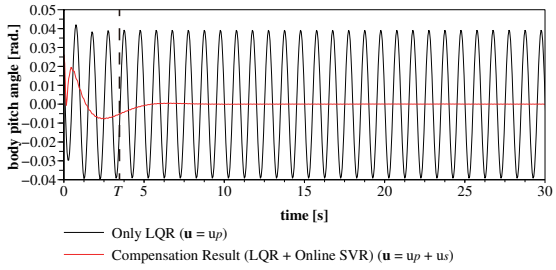


Figure 11. Control Response of the Body Pitch Angle ψ (1)

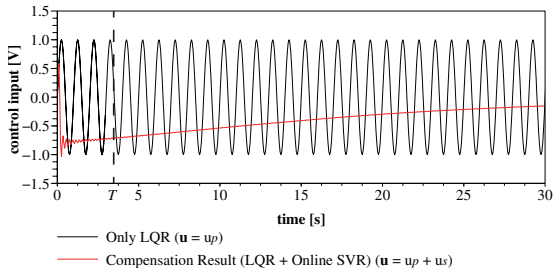


Figure 12. Control Response of the Control Input u (1)

method (are described in red solid line) are approaching to zero, with time. Next, we will focus on each result.

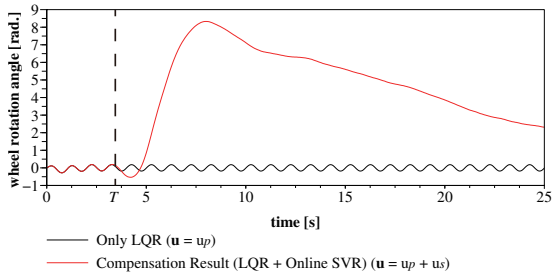


Figure 13. Control Response of the Wheel Rotation Angle θ (2) [4]

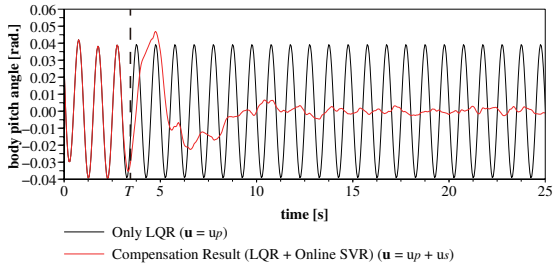


Figure 14. Control Response of the Body Pitch Angle ψ (2) [4]

Next, let's compare the results between proposed method and the results that showed in study [4]. Figure 13 through 15 show compensation results of the state of $\mathbf{x}_1$, and fig. 12 shows the compensation input using prediction result of $\mathbf{u}$, that used the method we proposed in a study [4]. Comparing each state and control inputs, it can be said that the convergence speed of experimental results is earlier than the former study.

Now let's focus on these results. In this study, the learning space and training sets will be changed "suddenly," when the period of the disturbance signal will be estimated. In this time, the current action and states are also suddenly changed by the compensation control input based on the former learning space and training sets. In a former study, training sets and learning space are including much prediction error, in phase of early training. Therefore, in the mechanism of State-action Pair Prediction, it will use former training sets, and will add new training sets to current learning space. Namely, results of prediction and revised action will be influenced by past prediction error. On the other hand, in the proposed method, the kernel matrix will ignore early learning results and early compensation result. In other words, it use results including less affection of prediction error, according to the time elapsed by the training set of out of the disturbance tendency period will not be used. Further, this system acquires the data each at the sampling times. Using these changed results, the proposed system derives an action that multiplied states to the optimal feedback gain for the future state. As a result, this system is stabilizing the inverted pendulum using current outside data and previous states and an action. As a result, we can be said that internal states and an action will converge to zero, according time course. From these viewpoints, we conclude the experimental results are reasonable.

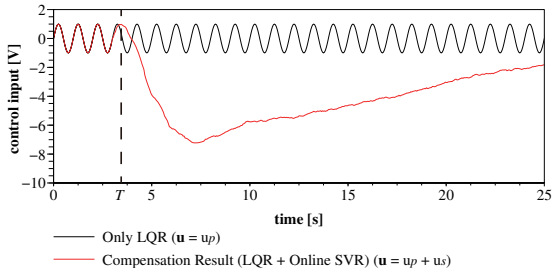


Figure 15. Control Response of the Control Input u (2) [4]

4 CONCLUSION

In this paper, the relationship between learning space and training sets for prediction and frequency of the disturbance signal that given by outside environment will be focused on. To achieve this problem, on the

basis of former our works, we proposed the method that reducing training sets and learning space, for prediction, based on prediction results that obtained by recent tendency of disturbance frequency, dynamically by using Nyquist-Shannon sampling theorem. Applying this proposed method, it was obtained that the body pitch angle of NXTway-GS was converged to zero, with time. In other words, the compensated action for rapid convergence was obtained, similarly as a normal training sets and learning space method.

From the verification experimental results, the proposed method could be converged to a desirable state as similar as fixed training sets. To be more specific, the slope of the body pitch angle of NXTway-GS will be converged to zero, based on state and action prediction and decision. Accordingly, the proposed method can be adapted to the situation of frequency property of disturbance will be concluded. From results of verification experiments, it can be concluded that the proposed system can predict what be defined by training sets and learning space that can be obtained the property of a disturbance signal, based on the Nyquist-Shannon Sampling Theorem. In addition, as a future work, we will confirm the response of the proposed system on actual robot.

REFERENCES

[1] P. Pivoňka, V. Veleba, M. Šeda, P. Ošmera, and R. Matoušek, "The Short Sampling Period in Adaptive Control," In Proceedings of the World Congress on Engineering and Computer Science 2009 Vol II WCECS, pp.724-729, 2009.

[2] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," Artificial Intelligence, Vol. 97 No. 1-2, pp.245-271, 1997.

[3] M. Sugimoto and K. Kurashige, "A Study of Effective Prediction Methods of the State-action Pair for Robot Control using Online SVR," Journal of Robotics and Mechatronics, Vol. 27 No. 5, pp.469-479, 2015.

[4] M. Sugimoto and K. Kurashige, "Real-time Sequentially Decision for Optimal Action using Prediction of the State-Action Pair," In Proceedings of 2014 International Symposium on Micro-NanoMechatronics and Human Science, pp.199-204, Nov.9-12, Nagoya, Japan, 2014.

[5] M. Sugimoto and K. Kurashige, "Future Motion Decisions using State-action Pair Predictions," International Journal of New Computer Architectures and their Applications, Vol. 5 No. 2, pp.79-93, 2015.

[6] R. Koggalage and S. Halgamuge, "Reducing the Number of Training Samples for Fast Support Vector Machine Classification," Neural Information Processing – Letters and Reviews, Vol. 2, No. 3, pp.57-65, 2004.

[7] G. Bontempi, Machine Learning Strategies for Time Series Prediction, Machine Learning Summer School, 2013. [Online]. Available: http://www.ulb.ac.be/di/map/gbonte/ftp/time_ser.pdf

[8] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, 2006.

[9] F. Parrella, Online Support Vector Regression. PhD thesis, Department of Information Science, University of Genoa, Italy, 2007.

[10] M. Sugimoto, H. Yoshimura, T. Abe, and I. Ohmura, "A Study on Model-Based Development of Embedded System using Scilab/Scicos," In Proceedings of the Japan Society for Precision Engineering 2010 Spring Meeting, Saitama, D82, pp.343-344, 2010.

[11] Y. Yamamoto, NXTway-GS Model-Based Design –Control of self-balancing two-wheeled robot built with LEGO Mindstorms NXT–. CYBERNET SYSTEMS CO.,LTD., 2009.

[12] R. Watanabe, Ryo's Holiday LEGO Mindstorms NXT, 2008.

[13] Y. Chang, C. Hsieh, K. Chang, M. Ringgaard, and C. Lin, "Training and testing low-degree polynomial data mappings via linear SVM," J. Machine Learning Research, vol.11, pp.1471-1490, 2010.

Controller Design based on Root Contour for Non-minimum Phase UAV System

Jia-Horng Yang*, Hua-Kai Hsu
Department of Electrical and Electronic Engineering,
Chung Cheng Institute of Technology, National Defense University,
No.75, Shihyuan Rd., Dasi Township, Taoyuan County 33551, Taiwan (R.O.C.)
yang.jiahorng@gmail.com*, withwind720721@gmail.com

ABSTRACT

This paper aims to analyze the scale model of the Cessna 182 and design the controller that meets the mission requirements. Different from other references that use Ziegler–Nichols method or another trial and error method to design the PID controller, this research applies the Root Contour method (RC) with multiple variables to certainly reflect the changes and relationships between the controller parameters and help design the better controller to improve the system performance and accuracy. In addition, face of the instability effect due to the non-minimum phase system from the transfer function between the speed and the elevator angle, this paper proposes a new design method of controller with the reverse multi-variable RC. The new method is practical and efficient to design the controllers and allow the operators can find the appropriate controller parameters more quickly and obtain the better system performances.

KEYWORDS

UAV, PID, Controller Design, Cessna 182, Non-Minimum Phase System

1 Introduction

Because the industrial technology and the progress speed of computer chip grow obviously, in addition, followed by the rise of the humanitarianism, the development of Unmanned Aerial Vehicle (UAV) is mellowed in recent years whether in the military use or the private technology. The research of UAV has been more and more capital into, e.g., disaster relief, military use, environmental explores, landscape filming, cross-border transportation and so on.

However, UAV has many limiting conditions, because there are too many uncertainties while flying (e.g., airflow disturbance, air density, temperature difference, compressive strength of carrier and sensor). Uncertain factors inhibit the applications of UAV.

In consideration of the difficulties of UAV's operating environment, this paper proposes an application combining RC and PID controller. This combination improves both the transient response and steady-state error performance of the UAV system effectively. Besides, the new design method of the reverse multi-variable RC can solve the problems from the non-minimum phase system and help to find the appropriate controller parameters more quickly to obtain the better system performances.

In addition to the improvement of the system behavior, the designing methods of this research not only extend the applications of UAV but also can also be the reference or applications for the follow-up researchers.

2 UAV Structure

2.1 Selection of the Carrier Configuration

In this paper, the scale model of the Cessna 182 is selected for the experimental UAV. And we call this UAV as "scale Cessna 182" for the clear definition to distinguish from the prototype Cessna 182. The ratio of the scale Cessna 182 to the prototype Cessna 182 is 1: 6.65.

Cessna series are manufactured for the aircrafts in Wichita, Kansas, USA, and they are widely used by the US Air Force for

education and training. So there are also many different scale patterns of prototype on the market. Because the performance of the scale machine is similar to the prototype, therefore, the scale model is often used to carry out the relevant experiments of the real flight equipment. The scale Cessna 182 is shown as Figure 1 [1], and the various parts of the body are seen in Figure 2 [2,3]. The geometric parameters of the scale Cessna 182 are shown in Figure 3 [4] and Table 1.

Figure 1. Scale Cessna 182 – 1/6.65 ratio to the Prototype Cessna 182 [1]

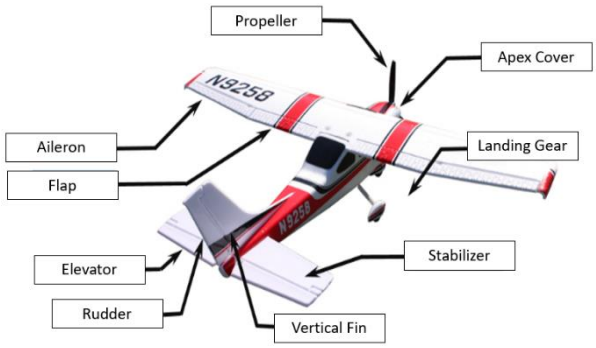


Figure 2. Various parts of Scale Cessna 182 [2,3]

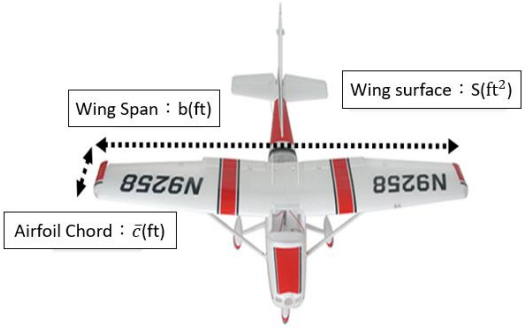


Figure 3. Geometry parameter definition of Scale Cessna 182 [4]

Table1. The geometry parameter of Scale Cessna182

	Scale factor ($SF = 6.65$)
Wing span (ft) b	5.4135
Wing surface (ft^2) S	3.9346
True airspeed (ft/sec) V_{P_1}	85.3511
Dynamic pressure (lbs/ft^2) $\bar{q}_1=\frac{1}{2}\rho V_{P_1}{}^2$	7.4586
Steady state angle of attack (deg) α	0

2.2 Model Constructing of the UAV

The basic motion of the aircraft is six degrees of freedom, that is, the horizontal movement along the X, Y, and Z, axes and the rotation around the X, Y, and Z axes, respectively (seen as Figure 4 [5]), which can be used to analyze the longitudinal and horizontal motions of the UAV movement. In this paper, the experimental UAV, scale Cessna 182, has the scale ratio 1: 6.65 of the prototype Cessna 182, and they have the similar system performances. Besides, since the basic dynamic and the derivation process of the aircraft model are general knowledge, therefore, due to page limit, this paper only illustrates a brief description of UAV model, and highlight the system analysis and controller design. And the detailed information for the derivation process of the UAV mathematical model can be referred to [6] or found on the webs.

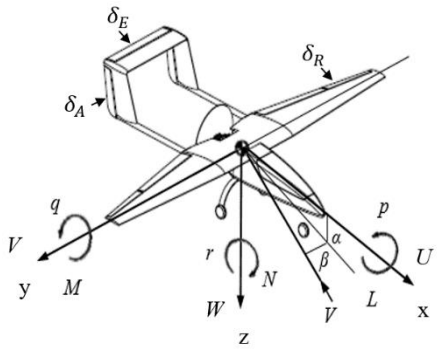


Figure 4. Six degrees of freedom for aircraft [5]

2.3 The Longitudinal Motion Equations of Scale Cessna 182

The following is a brief derivation and description of the longitudinal motion

equation for aircraft. Again, the complete derivation process of aircraft motion equation can be referred to the reference [6]. As we know, the longitudinal linearization equation of the aircraft can be also represented as equation (1).

$$\begin{aligned}\dot{u} &= -g\cos\Theta_1\theta + \left(X_u + X_{T_u}\right)u + X_\alpha\alpha + X_{\delta_E}\delta_E \\ V_{P_1}\dot{\alpha} &= -g\sin\Theta + Z_u u + Z_{\dot{\alpha}}\dot{\alpha} + Z_\alpha\alpha + \left(Z_q + V_{P_1}\right)\dot{\theta} + Z_{\delta_E}\delta_E \\ \ddot{\theta} &= \left(M_u + M_{T_u}\right)u + \left(M_\alpha + M_{T_\alpha}\right)\alpha + M_{\dot{\alpha}}\dot{\alpha} + M_q\dot{\theta} + M_{\delta_E}\delta_E\end{aligned} \quad (1)$$

where $\dot{u}$ is the derivative of the forward speed U, and $\dot{\alpha}$ is the derivative of the angle of attack (AOA), α. $\dot{\theta}$ and $\ddot{\theta}$ is the first and second derivative of the pitch angle θ. And the other parameter values are shown in Table 2 [6-8], Table 3 [7,8], and the Figure 4.

Table 2. The motion symbol of aircraft with six degree of freedom [6-8]

U	Forward speed	α	Angle of attack (AOA)
V	Slipping speed	θ	Pitch angle
W	Vertical speed	β	Angle of side slip (Sideslip angle)
p	Angular velocity of X axis	ϕ	Roll angle
q	Angular velocity of Y axis	ψ	Yaw angle
r	Angular velocity of Z axis	δ_E	Angle of elevator
L	Moment of X axis	δ_A	Angle of ailerons
M	Moment of Y axis	δ_R	Angle of rudder
N	Moment of Z axis	V_{P_1}	Indicated air speed
Θ_1	Pitch angle in the steady state	g	Gravity
$I_1 = \frac{I_{XZ}}{I_{XX}}$	Moment of inertia: I_1	$I_2 = \frac{I_{XZ}}{I_{ZZ}}$	Moment of inertia: I_2
SF		Scale factor: the ratio between UAV and real aircraft	

Table 3. Relationships for partial derivatives of longitudinal aerodynamics [7,8]

	Partial derivative of forward force	Partial derivative of normal force	Partial derivative of pitching moment M
Advance speed ($\boldsymbol{u}$)	X_u	Z_u	M_u
Angle of attack ($\boldsymbol{\alpha}$), AOA	X_α	Z_α	M_α
Angle of elevator ($\boldsymbol{\delta_E}$)	X_{δ_E}	Z_{δ_E}	M_{δ_E}
Vertical speed ($\boldsymbol{w}$)	X_w	Z_w	M_w
Pitch rate ($\boldsymbol{q}$)		Z_q	M_q
Change rate of vertical velocity ($\dot{\boldsymbol{w}}$)			$M_{\dot{w}}$
Rate of AOA($\dot{\boldsymbol{\alpha}}$)		$Z_{\dot{\alpha}}$	$M_{\dot{\alpha}}$
Thrust of engine to speed ($\boldsymbol{T_u}$)	X_{T_u}		M_{T_u}
Thrust of engine to AOA ($\boldsymbol{T_\alpha}$)			M_{T_α}

Using the Laplace transformation to the Equations (1) and assuming the initial conditions are zero; besides, the initial position of the system is balanced, then we can get the result as following:

$$\begin{aligned}&\mathcal{L}(\delta_E) = \delta_E(s) \\ &\mathcal{L}(\mathrm{u}) = \mathrm{u}(s)\ ;\ \mathcal{L}(\dot{u}) = \mathrm{su}(s)(\\ &\mathcal{L}(\alpha) = \alpha(s)\ ;\ \mathcal{L}(\dot{\alpha}) = \mathrm{s}\alpha(s) \\ &\mathcal{L}(\alpha) = \theta(s)\ ;\ \mathcal{L}(\dot{\theta}) = \mathrm{s}\theta(s)\ ;\ \mathcal{L}(\ddot{\theta}) = s^2\theta(s)\end{aligned}$$

Substitute the transformation result above into Eq. (1), we can get the Eq. (2):

$$\begin{aligned}&\left(\mathrm{s} - (X_u + X_{T_u})\right)\mathrm{u(s)} - X_\alpha\alpha(s) + X_\alpha\alpha(s) + g\cos\Theta 1(s) = X_{\delta_E}\delta_E(s) \\ &-Z_u u(s) + \left(s(V_{P_1} - Z_{\dot{\alpha}}) - Z_\alpha\right)\alpha(s) + \left(-\mathrm{s}(Z_q + V_{P_1}) + g\sin\Theta_1\right)\ \theta(s) = Z_{\delta_E}\delta_E \\ &-(M_u + M_{T_u})\mathrm{u(s)} - (M_{\dot{\alpha}}s + (M_\alpha + M_{T_\alpha}))\alpha(s) + s(s - M_q)\theta(s) = M_{\delta_E}\delta_E\end{aligned} \quad (2)$$

Let the angle of the elevator, $\delta_E(t)$, be the system input, and the output variables of X, Y, and Z axes are $u(t)$, $\alpha(t)$, $and\ \theta(t)$, respectively. So we can define the three transfer functions as $\left\{\frac{u(s)}{\delta_E(s)}, \frac{\alpha(s)}{\delta_E(s)}, \frac{\theta(s)}{\delta_E(s)}\right\}$ for prototype Cessna 182.

In this paper, for clearly distinguish the differences between the scale Cessna 182 and the prototype Cessna 182, we also add the symbol " s " to the parameters of prototype Cessna 182 to symbolize the scale Cessna 182. Therefore, in scale Cessna 182 system with the same assuming, if we set the angle of the elevator, $\delta_{E_s}(t)$, as the system input, and the output variables of X, Y, and Z axes are $u_s(t)$, $\alpha_s(t)$, $and\ \theta_s(t)$, respectively. So we can define the three transfer functions of the

scale Cessna 182 as $\left\{\frac{u_s(s)}{\delta_{E_s}(s)}, \frac{\alpha_s(s)}{\delta_{E_s}(s)}, \frac{\theta_s(s)}{\delta_{E_s}(s)}\right\}$.

In additions, before system analysis and controller design for the UAV (scale Cessna 182), the ratios of the aerodynamic parameters and geometric parameters between the scale Cessna 182 and prototype Cessna 182 must be considered. The ratio result is shown in Table 4 [8].

Since the ratio of scale Cessna 182 to prototype Cessna 182 is 1: 6.65, so we set scale factor, SF, be the value of 6.65. Therefore, we can obtain the partial derivative of longitudinal aerodynamics as Table 5 [8]. After reintegrating and deriving the Eq. (2) with the parameters in Table 5, we can get the three transfer functions of the scale Cessna 182 shown in Eq. (3~5) [8]:

Table 4. The ratios of longitudinal aerodynamics for Scale Cessna 182 [6,9]

	Scale factor *SF*=6.65		Scale factor *SF*=6.65
X_u/X_{u_S}	$\sqrt{SF}$	X_δ/X_{δ_S}	1
Z_u/Z_{u_S}	$\sqrt{SF}$	Z_δ/Z_{δ_S}	1
M_u/M_{u_S}	$SF\sqrt{SF}$	M_δ/M_{δ_S}	SF
X_α/X_{α_S}	1	M_q/M_{q_S}	$\sqrt{SF}$
Z_α/Z_{α_S}	1	Z_q/Z_{q_S}	$\sqrt{SF}/SF$
M_α/M_{α_S}	SF	$M_{\dot{w}}/M_{\dot{w}_S}$	SF
X_w/X_{w_S}	$\sqrt{SF}$	$Z_{\dot{\alpha}}/Z_{\dot{\alpha}_S}$	$\sqrt{SF}/SF$
Z_w/Z_{w_S}	$\sqrt{SF}$	$M_{\dot{\alpha}}/M_{\dot{\alpha}_S}$	$SF\sqrt{\frac{1}{SF}}$
M_w/M_{w_S}	$SF\sqrt{SF}$	$X_{T_u}/X_{T_{u_S}}$	$\sqrt{SF}$
$M_{T_u}/M_{T_{u_S}}$	$SF\sqrt{SF}$	$M_{T_\alpha}/M_{T_{\alpha_S}}$	SF

Table 5. Partial derivatives of longitudinal aerodynamics for Scale Cessna 182 [8]

X_{u_S}	-0.07839	X_{w_S}	0.228	M_{q_S}	-11.1841
Z_{u_S}	-0.75274	Z_{w_S}	-5.4448	Z_{q_S}	-1.7613
M_{u_S}	0	M_{w_S}	-1.5005	$M_{\dot{w}_S}$	-0.0771
X_{α_S}	19.459	X_{δ_S}	0	$Z_{\dot{\alpha}_S}$	-0.7678
Z_{α_S}	-464.71	Z_{δ_S}	-44.985	$M_{\dot{\alpha}_S}$	-6.584
M_{α_S}	-128.079	M_{δ_S}	-234.419	$X_{T_{u_S}}$	-0.0392
$M_{T_{u_S}}$	*0*	$M_{T_{\alpha_S}}$	*0*		

Transfer function of the speed $(u_s(s))$ to the elevator angle $(\delta_{E_s}(s))$:

$$\frac{u_s(s)}{\delta_{E_s}(s)} = \frac{-875.3631s^2 - 195940s + 1012100}{86.1189s^4 + 1985.9478s^3 + 16150s^2 + 2082.5s + 945.7337} \quad (3)$$

Transfer function of AOA $(\alpha_s(s))$ to the elevator angle $(\delta_{E_s}(s))$:

$$\frac{\alpha_s(s)}{\delta_{E_s}(s)} = \frac{-44.985s^3 - 20103s^2 - 2363.5s + 1730.9}{86.1189s^4 + 1985.9478s^3 + 16150s^2 + 2082.5s + 945.7337} \quad (4)$$

Transfer function of the pitch angle $(\theta_s(s))$ to the elevator angle $(\delta_{E_s}(s))$:

$$\frac{\theta_s(s)}{\delta_{E_s}(s)} = \frac{-19893s^2 - 105510s - 15567}{86.1189s^4 + 1985.9478s^3 + 16150s^2 + 2082.5s + 945.7337} \quad (5)$$

2.4 Equations of Lateral Motion

The linearized equation of lateral motion for scale Cessna 182 after deriving can be seen as Eq. (6) [8]:

$$\begin{aligned} (V_{P_1}\dot{\beta} + V_{P_1}\dot{\psi}) &= g\phi + Y_\beta\beta + Y_{\dot{\phi}}\dot{\phi} + Y_{\dot{\psi}}\dot{\psi} + Y_{\delta_A}\delta_A + Y_{\delta_R}\delta_R \\ \ddot{\phi} - \frac{I_{XZ}}{I_{XX}}\ddot{\psi} &= L_\beta\beta + L_{\dot{\phi}}\dot{\phi} + L_{\dot{\psi}}\dot{\psi} + L_{\delta_A}\delta_A + L_{\delta_R}\delta_R \\ \ddot{\psi} - \frac{I_{XZ}}{I_{ZZ}}\ddot{\phi} &= N_\beta\beta + N_{\dot{\phi}}\dot{\phi} + N_{\dot{\psi}}\dot{\psi} + N_{\delta_A}\delta_A + N_{\delta_R}\delta_R \end{aligned} \quad (6)$$

Where $\dot{\beta}$ is the differential of side slip angle, β. $\dot{\Psi}$ and $\ddot{\Psi}$ is the first and second differential of yaw angle, Ψ. And $\dot{\phi}$ and $\ddot{\phi}$ is also the first and second differential of roll angle, ϕ, respectively.The other parameter values can be seen in as Table 2, Table 6 [8], and Figure 4.

Table 6. Relationships for partial derivatives of lateral aerodynamics [8]

	Side force of *Y* axis	Slip moment of *L*-*X* axis	Yaw moment of *N*-*Z* axis
Sideslip angle ($\boldsymbol{\beta}$)	Y_β	L_β	N_β
Speed of roll angle ($\boldsymbol{p}$)	Y_p	L_p	N_p
Speed of yaw angle ($\boldsymbol{r}$)	Y_r	L_r	N_r
Ailerons angle ($\boldsymbol{\delta_A}$)	Y_{δ_A}	L_{δ_A}	N_{δ_A}
Rudder angle ($\boldsymbol{\delta_R}$)	Y_{δ_R}	L_{δ_R}	N_{δ_R}

In addition, the change rate of Euler angle and rotational angular velocity of the UAV with the small interference can be seen the same, so we can obtain the following result:

$$Y_{\dot{\phi}} = Y_P; Y_{\dot{\psi}} = Y_r; L_{\dot{\phi}} = L_p; L_{\dot{\psi}} = L_r; N_{\dot{\phi}} = N_p; N_{\dot{\psi}} = N_r$$

Using the Laplace transformation to the

Equations (6) and assuming the initial conditions are zero; besides, the initial flying state of the system is stable, then we can get the result as following:

$$\mathcal{L}(\delta_A) = \delta_A(s); \mathcal{L}(\delta_R) = \delta_R(s)$$
$$\mathcal{L}(\beta) = \beta(s); \mathcal{L}(\dot{\beta}) = s\beta(s)$$
$$\mathcal{L}(\phi) = \phi(s); \mathcal{L}(\dot{\phi}) = \mathcal{L}(p) = s\phi(s); \mathcal{L}(\ddot{\phi}) = \mathcal{L}(\dot{p}) = s^2\phi(s)$$
$$\mathcal{L}(\psi) = \psi(s); \mathcal{L}(\dot{\psi}) = \mathcal{L}(r) = s\psi(s); \mathcal{L}(\ddot{\psi}) = \mathcal{L}(\dot{r}) = s^2\psi(s)$$

Set $I_1 = \frac{I_{XZ}}{I_{XX}}, I_2 = \frac{I_{XZ}}{I_{ZZ}}$, then substitute I_1 and I_2 to the Eq. (6), we can get Eq. (7) as the following:

$$\begin{gathered}(sV_{P1} - Y_\beta)\beta(s) - (sY_p + gcos\Theta_1)\phi(s) + s(V_{P1} - Y_r)\psi(s) = Y_\delta\delta(s) \\ -L_\beta\beta(s) + s(s - L_p)\phi(s) - s(sI_1 + L_r)\psi(s) = L_\delta\delta(s) \\ -N_\beta\beta(s) - s(sI_2 + N_p)\phi(s) + s(s - N_r)\psi(s) = N_\delta\delta(s)\end{gathered} \quad (7)$$

In the aircraft motion, let the ailerons angle $\delta_A(t)$ and rudder angle $\delta_R(t)$ to be the input of the lateral motion. Therefore, the transfer functions of the lateral motion are $\left\{\frac{\beta(s)}{\delta_R(s)}, \frac{\phi(s)}{\delta_A(s)}, \frac{\psi(s)}{\delta_R(s)}\right\}$. Consider the scale factor between the scale Cessna 182 and the prototype Cessna 182 [2-4] and the partial derivatives of lateral aerodynamics [1,7], we can get the partial derivative value in proportion as Table 7 and Table 8 [8]. After reintegrating and deriving the Eq. (7) with the parameters in Table 6~8, we can also obtain the three transfer functions of the lateral motion for scale Cessna 182 (shown as Eq. (8~10) [8]):

Table 7. The ratio of the lateral motion for Scale Cessna 182

	Scale factor *SF*=6.65		Scale factor *SF*=6.65
Y_β/Y_{β_S}	1	$L_{\delta_A}/L_{\delta_{A_S}}$	*SF*
L_β/L_{β_S}	*SF*	$N_{\delta_A}/N_{\delta_{A_S}}$	*SF*
N_β/N_{β_S}	*SF*	$Y_{\delta_R}/Y_{\delta_{R_S}}$	1
Y_p/Y_{p_S}	$\sqrt{1/SF}$	$L_{\delta_R}/L_{\delta_{R_S}}$	*SF*
L_p/L_{p_S}	$1/\sqrt{1/SF}$	$N_{\delta_R}/N_{\delta_{R_S}}$	*SF*
N_p/N_{p_S}	$1/\sqrt{1/SF}$	L_r/L_{r_S}	$1/\sqrt{1/SF}$
Y_r/Y_{r_S}	$\sqrt{1/SF}$	N_r/N_{r_S}	$1/\sqrt{1/SF}$

Table 8. Partial derivatives of lateral motion for Scale Cessna 182 [8]

Y_{β_S}	-41.11	Y_{r_S}	0.71	$Y_{\delta_{R_S}}$	19.56
L_{β_S}	-201.163	L_{r_S}	-5.5185	$L_{\delta_{R_S}}$	32.053
N_{β_S}	61.6455	N_{r_S}	-3.1203	$N_{\delta_{R_S}}$	-67..7635
Y_{p_S}	-0.249	$Y_{\delta_{A_S}}$	0	N_{p_S}	-0.9284
L_{p_S}	-33.4465	$L_{\delta_{A_S}}$	499.149	$N_{\delta_{A_S}}$	-22.6765

Transfer function of the sideslip angle ($\beta_s(s)$) to the rudder angle ($\delta_{R_s}(s)$):

$$\frac{\beta_s(s)}{\delta_{R_S}(s)} = \frac{19.56s^4 + 10483.9816s^3 + 1387923.0948s^2 - 17870.931s}{85.3511s^5 + 20795.8516s^4 + 428413.4354s^3 + 1466182.8401s^2 + 18755.221s} \quad (8)$$

Transfer function of roll angle ($\phi_s(s)$) to ailerons angle ($\delta_{A_s}(s)$):

$$\frac{\phi_s(s)}{\delta_{A_S}(s)} = \frac{42602.9162s^3 + 833502.2803s^2 + 2609905.1943s}{85.3511s^5 + 20795.8516s^4 + 428413.4354s^3 + 1466182.8401s^2 + 18755.221s} \quad (9)$$

Transfer function of yaw angle ($\psi_s(s)$) to rudder angle ($\delta_{R_s}(s)$):

$$\frac{\psi_s(s)}{\delta_{R_S}(s)} = \frac{-5783.6893s^3 - 1304871.8347s^2 - 332356.8248s - 114340}{85.3511s^5 + 20795.8516s^4 + 428413.4354s^3 + 1466182.8401s^2 + 18755.221s} \quad (10)$$

2.5 Transfer Function of the Actuator

When analyzing the aircraft system, in reality, we should consider the actuator. Generally, the actuator is a first order system. Its transfer function can be represented as Eq. (11), the parameters of the actuator is shown as Table 9.

$$\frac{\delta}{v} = \frac{K_a}{\tau s + 1} \quad (11)$$

Table 9. Parameter of actuator.

δ_E	Angle of elevator
v_E	Input voltage
K_a	Gain value of servo motor
τ	Time constant of servo motor

In general, the initial gain value of servo motor is set as 1, and time constant of servo

motor is usually between 0.05 ~0.25 second. In this paper, we set the time constant of servo motor is 0.1 second.

Besides, the up direction with respect to the elevator surface is negative in definition. The up direction with respect to the rudder surface and ailerons surface is on the contrary. Hence, the transfer functions of servo motors for elevator, rudder, and ailerons can be defined as Eq. (12):

$$\begin{cases} \text{For Elevator}: \dfrac{\delta_E}{v_E} = -\dfrac{10}{s+10} \\ \text{For Rudder}: \dfrac{\delta_R}{v_R} = -\dfrac{10}{s+10} \\ \text{For Ailerons}: \dfrac{\delta_A}{v_A} = \dfrac{10}{s+10} \end{cases} \tag{12}$$

3 Research Methods

This study is based on the root locus method, and combines with PID theory to design a robust controller. Different from other references that use Ziegler–Nichols method or another trial and error method to obtain the PID parameters, the controller design in this paper is more flexible and adjustable in response to the environment and requirement.

For controller design in this paper, firstly, applying the root locus method to find the change of the single variable, secondly, uses the root contours method to solve the coupling effects among the multiple controller parameters and find the optimal parameter of controller further. Moreover, we propose a new design method of controller with the reverse multi-variable RC. The new method is practical and efficient to design the controllers. This new method can overcome the negative effect due to the non-minimum phase system that Ziegler–Nichols method cannot deal with. Even though there are already some methods, e.g., fuzzy PID, sliding mode control, multi-model adaptive control, and so on [10-13]. However, these methods are rarely used to help to stabilize the aircraft s or applied to avoid the obstacle, and perform some tasks. The new method in this paper, on the contrary; really improves the both transient response and steady-state performance. It can also overcome the difficulties of the controller design due to the high-order aircraft systems. Besides, it helps to design the suitable controllers to meet the task requirements and extend the UAV applications. The following are some methods of controller design.

3.1 PID Controller

The advantages of orthodox PID controller include: (1) PID is theoretical; (2) easy to achieve; (3) high control accuracy. However, the controller designers must deeply understand the system characteristics and the operating environment. Besides, these PID controllers have to be designed via root locus criterion, and then the operators can design the suitable PID controllers. However, many operators design PID controllers based on Ziegler–Nichols method now. Therefore, these PID controllers lack the advantages discussed above from orthodox PID. The greatest strengths of these PID based on Ziegler–Nichols method are that it is easy to obtain the PID parameters, even though theses parameters are not very appropriate, but these parameters did help to obtain some slightly better system performance; besides, the operators do not have to learn additional skills of controller design, e.g., root locus (RL), root contour (RC), and so on. However, Ziegler–Nichols method really lacks the flexibility and accuracy. It is not easy to adjust some specific parameters to meet some special demands or achieve some special accuracy in real time. It's just a kind of trial and error method and has to try many times to obtain a bit better parameter. This is also the main reason why we didn't apply the easier, Ziegler–Nichols method to design PID or other controllers in this paper.

The general form of PID transfer function is shown as Eq. (13), where the related parameters can be seen in Table 10.

$$u_{PID}(t) = K_p e(t) + K_i \int_0^t e(\tau)dt + K_d \frac{de(t)}{dt} \tag{13}$$

Table 10. Parameter of PID controller

K_p	Proportional gain
K_i	Integral gain
K_d	Derivative gain
e	Error
t	Time
τ	Variable of integration

The ideal PID controller can be seen as the aggregation of two zeros and one pole which is located at origin. It belongs to an active circuit, which can be used to improve the system transient response and steady-state error at the same time.

Because the RL can just select one controller parameter at a time, therefore, when designing PID, it's usually separated into two steps: (1) PD controller improves the system transient response; and (2) PI controller improves steady state error of the system. These two steps are no prioritizations. In addition, using RL does help to find the optimal poles and zeros of the closed-loop system. The rule is that the total angel summation of poles and zeros located at the root locus has to be 180° (seen as Figure 5).

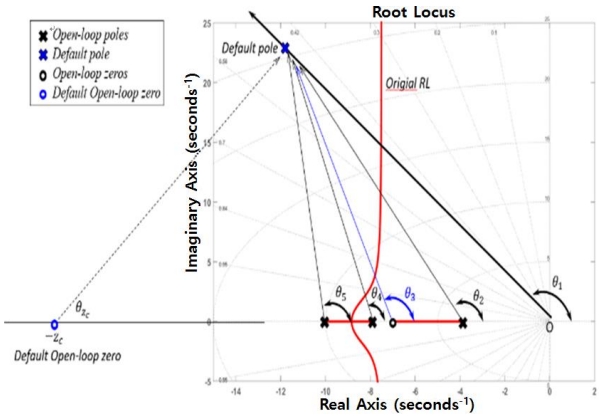


Figure 5. The total angel summation of poles and zeros located at the root locus has to be 180°.

According to our experience, after finding one parameter, we suggest to simulate before searching the next one. Besides, after obtaining the whole controller parameters, we still need to consider the root sensitive and the coupling effects among the parameters.

In additions, if we have to adjust the controller parameters to meet the new requirement of the tasks suddenly, in accordance with past work experience, the parameters adjustment of the PID controller can comply with the Table 11. This is why our controller design more flexible than Ziegler-Nichols method or other designs.

Table 11. Principles of PID controller design.

Para-meter	Rise time	Setting time	Over-shoot	e(∞)	Stabi-lity
↑ K_p	Decline	slightly increase	Increase	Decline	Decline
↑ K_i	Slightly decline	Increase	Increase	Suddenly decline	Decline
↑ K_d	Slightly decline	Decline	Decline	The same	Increase

3.2 Ziegler–Nichols Method

The Ziegler-Nichols method is created by two engineers of Taylor Instrument. The method is used to obtain the approximate value of PID parameters. The first step of Ziegler-Nichols method is to set both the integration gain K_i and differential gain K_d as " 0 ", then gradually increases the value of proportional gain K_p from zero to its maximum value K_u. At this time, the output of controller will oscillate by a constant value, and its oscillation period is T_u. Then using maximum value K_u and oscillation period T_u to obtain the PID controller parameter according to the types of controllers (seen as Table 12). Of course, the controller parameters will not be accurate; hence the users have to implement the trial and error method to revise the response. However, many PID designers use Ziegler-Nichols method to replace the orthodox PID method whose design process has to combine RL or RC. Because Ziegler-Nichols method is too easy, however, it is also too rough and not accurate compared to expected value. It can't also be adjusted the specific one or all controller parameters in real time to meet the requirement suddenly. This is the most significant disadvantage of the Ziegler-Nichols method. Besides, the Ziegler-Nichols method can't be used with non-minimum phase or unstable systems. Nevertheless, there are many coupling effects among the forces and moments of the aircrafts. There always exists non-minimum phase system that yields the system to diverge. In this case, Ziegler-

ISBN: 978-1-941968-42-0 ©2017 SDIWC

Nichols method can't solve the problem, so that's main purpose why we create a new design method based on RC to overcome the non-minimum phase problem in this paper.

Table 12. Ziegler–Nichols method

Controller	K_p	K_i	K_d
P	$0.5K_u$	0	0
PI	$0.45K_u$	$0.45\,K_u/0.83T_u$	0
PID	$0.60K_u$	$0.6K_u/0.5T_u$	$0.6K_u/0.125T_u$

3.3 Root Contours Method

The best assistant tool of PID controller is root locus method (RL). It can help designers to obtain the desired characteristic roots or estimate the trend of the future system responses by the change of single variable (e.g., gain or other variables). However, while the system is not only one variable, then we need another method in response to multiple variables. This design method is called as "root contour" (RC). It can deal with the design problems of the multiple variables. The following is the brief description of the RL and RC: (1) RL: if the open-loop system transfer function is shown as Eq. (14) and Eq. (15):

$$G(s)H(s) = KG_1(s)H_1(s) = \frac{KR_1(s)}{Q_1(s)} \tag{14}$$

$$G_1(s)H_1(s) = \frac{R_1(s)}{Q_1(s)} \tag{15}$$

where

$$\begin{cases} Q_1(s) = s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0 \\ R_1(s) = s^m + b_{m-1}s^{m-1} + \cdots + b_1 s + b_0 \\ a_1\,,\ a_2\,,\ \ldots\,,\ a_n\,,\ b_1\,,\ b_2\,,\ \ldots\,,\ b_n \text{ is fixed real number} \\ \text{K = is constant, range} -\infty < K < \infty \end{cases}$$

Both Eq. (14) and Eq. (15) are the general forms of RL. And assuming the closed-loop system transfer function is also shown as equation (16):

$$\frac{Y(s)}{R(s)} = \frac{G(s)}{1 + G(s)H(s)} \tag{16}$$

where $Y(s)$ is system output, $R(s)$ is the reference input. From equation 15, we can see that the system characteristic equation is shown in Eq. (17):

$$1 + G(s)H(s) = 0 \tag{17}$$

Since the characteristic polynomial equals to zero, therefore, we can get the result as Eq. (18). The same as Eq. (14) and Eq. (15), the Eq. (18) is also the general form of RL. And then we can use Eq. (18) to help obtain the controller parameters.

$$Q_1(s) + KR_1(s) = 0 \tag{18}$$

For example, if a characteristic equation of system is shown in Eq. (19):

$$s(s+1)(s+2) + K(s+1)s + (s+2)=0 \tag{19}$$

Dividing both side of Eq. (19) by the term that does not contain the variable K, then we can get Eq. (20):

$$1 + \frac{K(s+1)s}{s(s+1)(s+2) + +(s+2)} = 0 \tag{20}$$

Comparing the Eq. (14~18), hence, we can easily get the result as Eq. (21). Afterword, we use this RL general form to obtain the controller parameters.

$$\frac{R_1(s)}{Q_1(s)} = \frac{(s+1)s}{s(s+1)(s+2) + +(s+2)} \tag{21}$$

So we can use root locus method to solve the system too.

RC is the extended criterion from RL to deal with the multiple-variables problem. The criterion can be summarized as following:

Since the characteristic polynomial is 0, assume a closed-loop system has two variables in the characteristic polynomial. Its RC general form is similar to Eq. (18) as the Eq. (22):

$$Q(s) + K_1R_1(s) + K_2R_2(s) = 0 \tag{22}$$

The First step of RC is to set any one of the variables is zero. In this description, we

set variable K_2 is 0, then Eq. (22) can be rewrite as Eq. (23):

$$Q(s) + K_1 R_1(s) = 0 \tag{23}$$

Then we divide both side of Eq. (23) by the term that does not contain the variable K_1, then we can get Eq. (24):

$$1 + \frac{K_1 R_1(s)}{Q(s)} = \frac{Q(s) + K_1 R_1(s)}{Q(s)} = 0 \tag{24}$$

Form Eq. (24), we can find the general form of root as transfer function: $\frac{R_1(s)}{Q(s)}$, then we draw the RL according the Eq. (24).

Secondly, in Eq. (22), we set K_1 is a random constant, and K_2 is the system variable now. We divide both side of Eq. (22) by the term that does not contain the variable K_2, then we can get Eq. (25) and Eq. (26):

$$1 + \frac{K_2 R_2(s)}{Q(s) + K_1 R_1(s)} = \frac{Q(s) + K_1 R_1(s) + K_2 R_2(s)}{Q(s) + K_1 R_1(s)} = 0 \tag{25}$$

$$G_2(s) H_2(s) = \frac{R_2(s)}{Q(s) + K_1 R_1(s)} \tag{26}$$

In Eq. (25) and Eq. (26), at this time, K_2 is a variable and K_1 is a random constant we chose. Now we can draw the RL of the Eq. (26).

In the final step of RC, we repeat the second step above several times (at least three times). So we continue to choose different constant as K_1 values, at the same time, K_2 is still a variable. We repeatedly draw the RL according to the Eq. (26). At last, we can find that after combing the tracks of the repeated RL, RL tracks will converge into a pattern. This pattern is RC. From RC, we can also find the trends of the closed-loop roots and system characteristics.

Different locations of poles and zeros have their distinct, various properties. We use this concept of RC and combine the method of controller design to place the poles and zeros of closed-loop at the desired locations, and then we will obtain the expected response.

4 Research results

4.1 Elevator Angle to Angle of Attack

The transfer function of AOA to the elevator angle as equation (4) is shown above, and the block diagram is shown as Figure 6. As mentioned above, most of users apply the Ziegler – Nichols method to design PID controller. Reference [8] is an example, too. We repeat reference [8] and apply Ziegler – Nichols method and Table 12 to design the PID controller. We can get three parameters ($K_p = 2$, $K_i = 15$, and $K_d = 0.29$) of Eq. (12). Besides, this controller equation can be also represented the form such as Eq. (27):

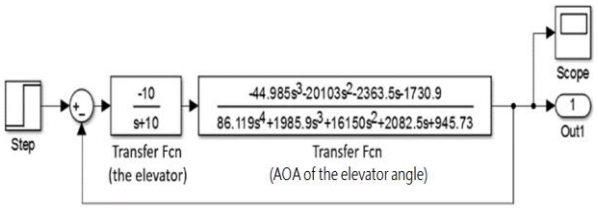


Figure 6. Block diagram of elevator angle to AOA

$$C_1(s) = \frac{0.29 * (s^2 + 6.9s + 51.7)}{s} \tag{27}$$

Different from other controller designs based on the Ziegler–Nichols method, we apply the RC mentioned above to design the controller to highlight the superiority of the designed controller in this study. We assume that the form of PID controller is shown as Eq. (28). Specially, we set the system gain 0.29 in Eq. (28) to be the same as Eq. (27) based on the Ziegler–Nichols method. That will be a clear comparison between reference [8] and this study, good contrast between the Ziegler–Nichols method and our design based on RC combining with the orthodox PID criterion. Afterwards, the next procedure is to find the other parameters (K_1 and K_2) of the Eq. (28):

$$C_2(s) = \frac{0.29 * (s^2 + K_1 s + K_2)}{s} \tag{28}$$

As mentioned above, the first step is to set any one of the variables as zero. Therefore, we set $K_2 = 0$ and K_1 is a variable, and then we get the root trajectory (seen as Figure 7).

RC analysis for elevatotr angle to AOA

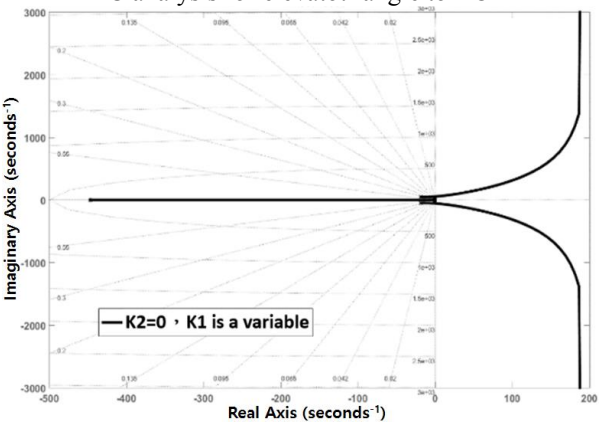


Figure 7. K_2= 0, K_1 is a variable

In the Figure 7, we zoom in the region near the origin point (seen as Figure 8) to easily figure out the root trajectory and select three parameters to be candidates of variable K_1. One of these three K_1 parameters is 6.9 which is the same as reference [8] for comparison. $K_1 = 6.9$ selected by the Ziegler–Nichols method in reference [8] is not perfect enough, since this point is too close to jw-axis. We randomly pick up two points, $K_1 = 8.9$ and $K_1 = 11$. According to RL and RC, all system response of them are much better than $K_1 = 6.9$ taken by Ziegler–Nichols method in reference [8]. Besides, from the Figure 8, it can be seen that $K_1 = 11$ is the leftmost half of root trajectory. Therefore, its transient response is the best. The transient response of $K_1 = 6.9$ is the worst compared to other values. Table 13 represents the three responses from the different K_1. In these cases, as mentioned above, we assumed $K_2 = 0$ and K_1 is a variable.

RL for elevator angle to AOA with $K_2 = 0$, K_1 is a variable

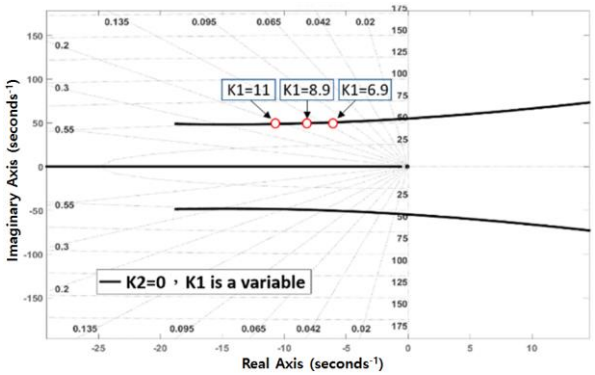


Figure 8. Partial enlargement around the origin point of Figure. 7

Table 13. Step responses from different K_1, where $K_2 = 0$.

	Peak time (sec)	Rise time (sec)	Setting time (sec)	Steady state error	Over-shoot (OS%)
$K_1 = 6.9$	11.4	0.349	3.61	0.073	1.65
$K_1 = 8.9$	8	0.265	2.63	0.058	0.751
$K_1 = 11$	none	0.192	0.6	0.047	none

In additions, in the reference [8], the authors designed the parameter $K_1 = 6.9$ of their PID controller. Although we can design many different values of K_1 whose responses are all better than $K_1 = 6.9$ in reference [8]. However, the better transient response we want, the expensive equipment we have to pay. Therefore, to avoid taking too extreme value of K_1, in this paper, we take the better response than $K_1 = 6.9$, but not too extreme. Finally, we choose $K_1 = 8.9$. Of course, we can also adjust K_1 to other values according to the hardware specifications or task requirement in the future.

Continuing to design PID controller, the second step, K_2 is a variable, and we set $K_1 = 8.9$ for a constant value. Then we can get the root contours shown in Figure 9. Again, we enlarge the region near the origin point of the Figure 9 to analyze the characteristic roots (seen as Figure 10). Since K_2 was selected as 51.7 in reference [8], we randomly choose another two parameters ($K_2 = 41$ and 62) around $K_2 = 51.7$. In Figure 10, we can find that the damping ratio of $K_2 = 62$ is close to 0.707. Therefore,

$K_2 = 62$ will have better transient response, including the rising time, setting time, and peaking time. However, $K_2 = 62$ has less damping ratio compared to $K_2 = 41$ and $K_2 = 51.7$. Hence, according to the property of damping ratio and RC criterion, we can predict the system with controller parameter $K_2 = 62$ in Eq. (28) will have larger overshoot than another two systems. Operators can choose the optimal K_2 value according the task requirement. The step responses of the three different K_2 can be seen as Table 14.

RC analysis for elevator angle to AOA

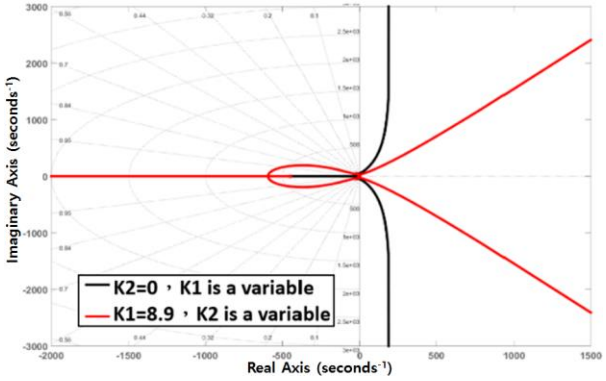


Figure 9. The RC： K_1= 8.9 , K_2 is a variable

RC analysis for elevator angle to AOA

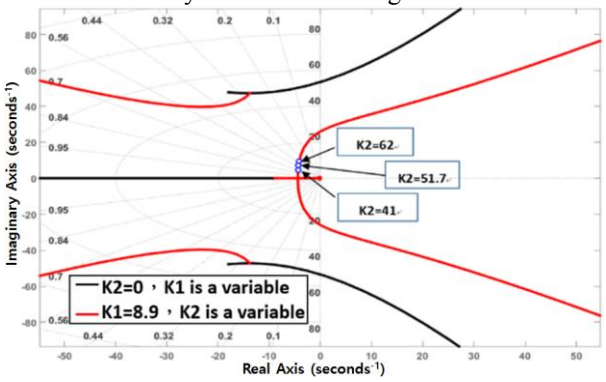


Figure 10. Partial enlargement around the orginal point

Table 14. Step responses from different K_2, where $K_1 = 8.9$

	Peak time (sec)	Rise time (sec)	Setting time (sec)	Steady state error	Over-shoot (OS%)
$K_2 = 41$	0.721	0.338	0.977	0	4.6
$K_2 = 51.7$	0.577	0.266	0.868	0	8.57
$K_2 = 62$	0.506	0.222	0.775	0	12.2

In summary, it is possible to a better controller for the transfer function of the elevator angle to AOA. The controller transfer function in this paper is shown as Eq. (29):

$$C_2(s) = \frac{0.29 * (s^2 + 8.9s + 41)}{s} \tag{29}$$

Figure 11 and Table 15 are the responses of reference [8] (Original PID) and our design (Designed PID). We can find that the response of the Designed PID based on RC criterion is better than the controller in reference [8] (Original PID). Comparing the controller design based on the Ziegler-Nichols method in reference [8], our controller design combines the orthodox PID criterion and RC. Not only shorten the setting time of the system, but also reduce the system overshoot. It is evident that the method of controller design in this paper is effective for aircraft and UAV.

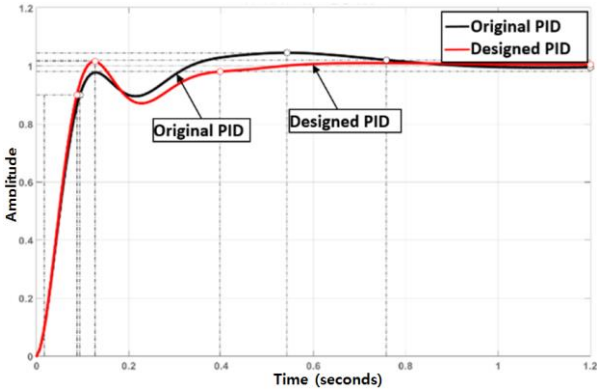


Figure 11. Closed-loop responses of reference [8] (Original PID) and our design (Designed PID)

Table 15. The comparison of reference [8] (Original PID) and our design (Designed PID)

	Peak time (sec)	Rise time (sec)	Setting time (sec)	Steady state error	Over-shoot (OS%)
Original PID	0.1252	0.07	0.7552	0	7.2808
Designed PID	0.127	0.07	0.397	0	1.5

4.2 Speed to Elevator Angle

The transfer function of speed to the elevator angle is as equation (3) shown above, and the block diagram is shown as Figure 12.

ISBN: 978-1-941968-42-0 ©2017 SDIWC

As mentioned above, most of users apply the Ziegler–Nichols method to design PID controller. Reference [8] is an example, too. However, the authors in reference [8] indicated that the transfer function of speed to the elevator angle (as Eq. (3) and Figure 12) is unstable. The closed-loop response of Scale Cessna 182 is divergent. And Ziegler–Nichols method could not help stabilize the system.

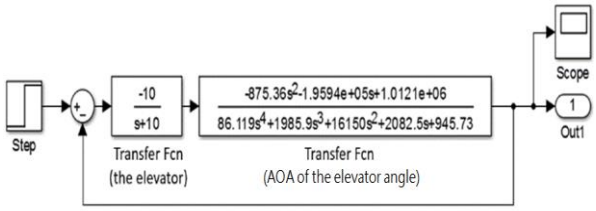


Figure 12. Block diagram of speed to elevator angle

From Figure 12, after analyzing the distribution of poles and zeros in the system, we can find that there is a zero located at right hand side (RHS) of the s-plane. Therefore, this is a non-minimum phase system. In additions, the step response of this closed-loop system is divergent (seen as Figure 13). Since the system is unstable, therefore, it is impossible to find maximum gain K_u and oscillation frequency T_u. Hence, the Ziegler-Nichols method is useless to help design the PID controller to stabilize the system of speed to elevator angle.

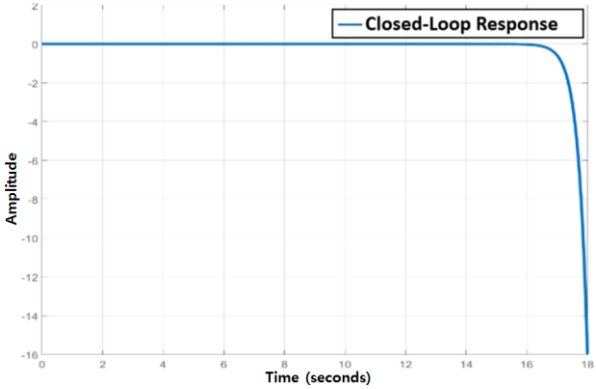


Figure 13. Closed-loop response: speed to elevator angle

According to the unstable situation above, in this paper, we do not use Ziegler-Nichols method and change to apply RC criterion to help solving this problem. However, from the Figure 14 and Figure 15 we find that, there are too many characteristic roots located at RHS of s-plane. Therefore, even though we can design a controller to overcome this case, the procedures are complicated, and sometimes, the controller parameters are maybe not perfect. The first consideration in the case is how to haul the roots trajectory to the left hand side (LHS) of the s-plane. In view of this, we try to create a new method to design the controller based on RC criterion. The transfer function of this new controller is shown as Eq. (30), and we define this new controller as "Reverse Gain PID Controller". Equation (30) has three variables. Comparing to Eq. (28), the additional gain $(-K_1)$ of Eq. (30) is designed to draw the most root trajectory in RHS of the Figure 14 and Figure 15 to the LHS.

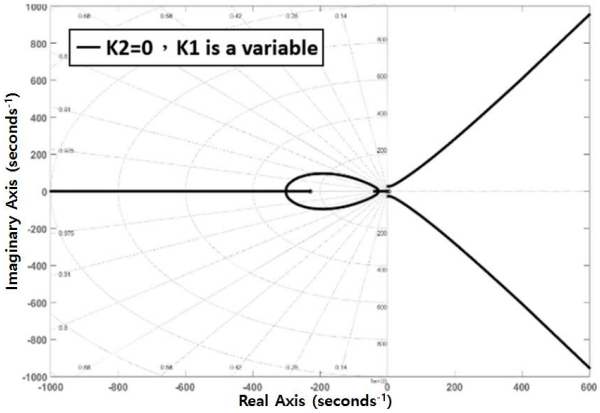


Figure 14. RL and RC analysis of speed to elevator angle

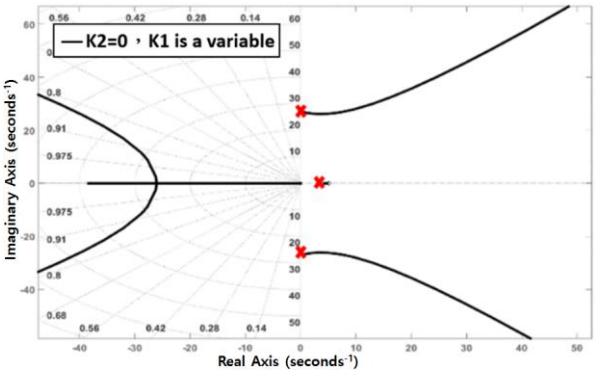


Figure 15. Partial enlargement around the origin point of Figure 14.

$$C_3(s) = \frac{-K_{1*}(s^2 + K_2 s + K_3)}{s} \tag{30}$$

As mentioned above, the first step of design procedures in Eq. (30) is to set $K_2 = K_3 = 0$ and K_1 is a variable, then we can get the root trajectory shown as Figure 16.

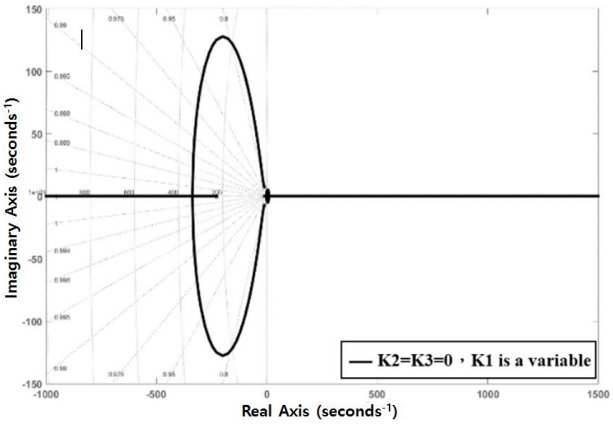


Figure 16. RC analysis for speed to elevator angle with different K_1 values in Eq. (30)

In the Figure 16, we zoom in the region near the origin point (seen as Figure 17(a),(b)) to easily figure out the root trajectory. We randomly select three parameters to be candidates of variables K_1, including 0.0092, 0.01805, and 0.03146. In Figure 17(a), we find that when $K_1 = 0.01805$, the damping ratio of the system is close to 0.707, and its overshoot will be better than the one of $K_1 = 0.0092$. In additions, even though $K_1 = 0.03146$ seems has less overshoot than the one of $K_1 = 0.01805$, however, $K_1 = 0.03146$ also lead to the system with a very small characteristic root close to origin (seen as the Figure 17(b)), and this small characteristic roots will cause a large setting time. Therefore, for fast response, we can choose K_1 equals to 0.0092 or 0.03146. In this case, integrating the performance of the damping ratio and speed of the response, we choose 0.01805 as the K_1 value.

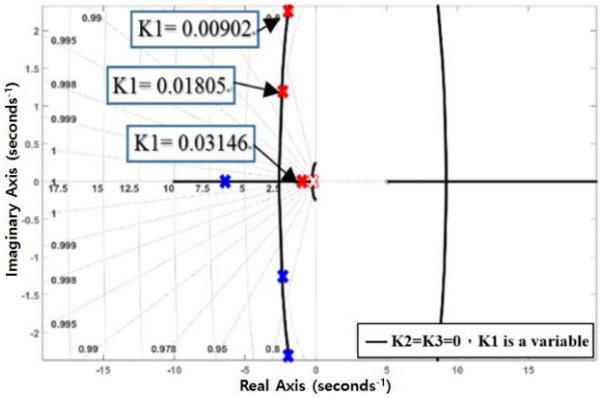


Figure 17(a). Partial enlargement of Figure 16

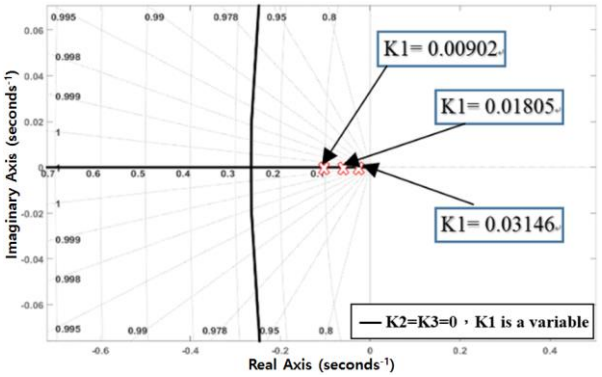


Figure 17(b). Partial enlargement around the origin point of Figure 16.

The second step of the controller design for the non-minimum phase system is to set K_3=0, $K_1 = 0.01805$, and K_2 is a variable. The RC is shown as Figure 18. We arbitrarily pick three points, and the step responses of these three characteristic roots can be seen in Table 16. Integrating the system performances of the different K_2 valus, finally, we select K_2 = 0.495.

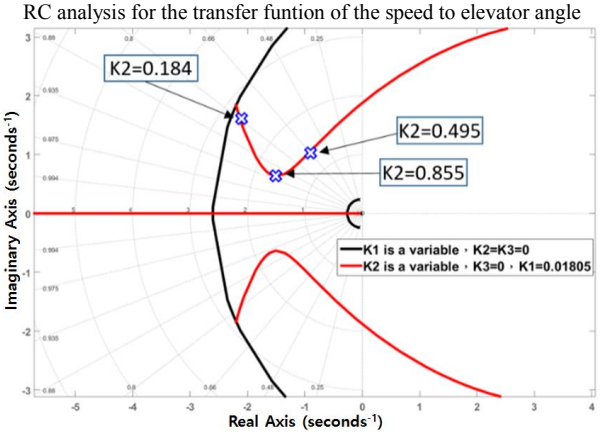


Figure 18. RC analysis with K_3=0, $K_1 = 0.01805$, and K_2 is a variable in Eq. (30)

Table 16. K_3= 0, K_1 = 0.01805, and K_2 is a variable. 0 < K_2< 2.2 for stability

	Peak time (sec)	Rise time (sec)	Setting time (sec)	Steady state error	Over-shoot (OS%)
K_2 = 0.184	None	8.54	16.1	0.22	None
K_2 = 0.495	None	2.37	4.49	0.09	None
K_2 = 0.855	3.09	1.17	6.27	0.06	17.3

The third step, we let K_3 is the variable, K_2 = 0.495, and $K_1 = 0.01805$ are the constants. The RC of this setting is shown as Figure 19. We arbitrarily pick three points, and the step responses of these three characteristic roots are shown in the Table 17.

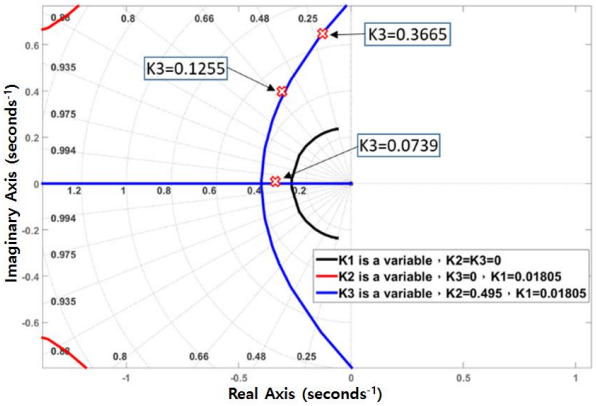


Figure 19. RC analysis with $K_1 = 0.01805$, $K_2 = 0.495$, and K_3 is a variable in Eq. (30)

Table 17. K_1 = 0.01805, $K_2 = 0.495$, and K_3 is a variable. 0 < K_3< 0.56 for stability

	Peak time (sec)	Rise time (sec)	Setting time (sec)	Steady state error	Over-shoot (OS%)
K_3 = 0.0739	None	11.2	21.3	0	None
K_3 = 0.1255	11.3	5.03	14	0	3.27
K_3 = 0.3665	5.72	1.83	34.9	0	57.6

Integrating the system performances of the different K_3 valus, we can find that when $K_3 = 0.1255$, as the whole, the system will have more reasonable, suitable response for UAV. Therefore, in this paper, the new design PID controller for transfer function of speed to the elevator angle is represented as Eq. (31). This design is useful to help to overcome the negative effects of the non-minimum phase system. The comparison of the system responses between the original system [8] and our design ($C_4(s)$) is also shown in Table 18.

$$C_4(s) = \frac{-0.01805 * (s^2 + 0.495s + 0.1255)}{s} \quad (31)$$

Table 18. Closed-loop responses based on the transfer function of the speed to the elevator angle

	Peak time (sec)	Rise time (sec)	Setting time (sec)	Steady state error	Over-shoot (OS%)
Original system [8]	System is divergent				
$C_4(s)$	2.17	0.631	11.4	0	31.8

5 Conclusion

The main objective of this paper is to analyze the UAV system, scale Cessna 182, and design a suitable PID controller according to task requirements. Different from Ziegler–Nichols method and other common trial and error method, the controller design in this paper combines the RC and orthodox PID criterion that help operators to design the more suitable controller. In additions, the PID controller based on RC is more flexible to face the sudden difficulties or demands.

Besides, RC with multiple variables of controller provides the relationship between the controller parameters. Moreover, it does reflect the effects of the parameter changes and help design the reasonable controller to improve the response performance and accuracy of the system. In this paper, the controller of AOA to the elevator angle is the best validation.

Furthermore, due to the non-minimum phase system of speed to elevator angle in scale Cessna 182, Ziegler–Nichols method can't solve the problem [8]. In this paper, we create a new design method called "Reverse Gain PID Controller" to help overcome the negative effect from non-minimum phase, and stabilize the divergent system. In additions, the new method in this research is academic and efficient that can be the reference or applications for the follow-up researchers.

6 Acknowledgement

This research can be accomplished all thanks for the supports from Ministry of Science and Technology, MOST of ROC. (MOST 105-2221-E-606-013) that makes the research completed smoothly.

7 Reference

[1] http://phoenixmodel.com/Product.aspx?ProductId=51 (2017.3.15)
[2] http://www.dianliwenmi.com/postimg_15172960.html (2015.09.03)
[3] http://ep.yimg.com/ay/yhst-10339770732811/cessna-182-v3-pro-series-rc-plane-15.gif (2015.09.03)
[4] https://www.aliexpress.com/item-img/Free-shipping-RC-airplane-cessna-182-EPO-foam-frame-kit-aeromodeling-remote-control-rc-plane/32495444238.html# (2017.05.08)
[5] F.-B. Hsiao and C.-S. Lee, *The Realization of Optimal Stability Augmentation Autopilot for Unmanned Air Vehicle*, Taiwan Tainan, National Cheng Kung University, pp.10-12, 2008.
[6] S. Bogos and I. Store, “Similarity Criteria for "full" and "scale" Aircraft on the Lateral Stability Analysis”, *U.P.B. Sci. Bull.*, Serial D, Vol. 74, Iss. 4, pp.14-26, 2012.
[7] M. R. Napolitano, *Aircraft Dynamics: From Modeling to Simulation*, Virginia, John Wiley & Sons, Inc., pp.352-375, 2012.
[8] C.-Y. Yang, *On the Modeling and Stability Augmentation of a Scale Unmanned Aerial Vehicle*, Taiwan Taichung, Feng Chia University, 2013.
[9] W. W. Durgin and C.-W. Kim, “Scale Modeling of Cessna 172”, American Institute of Aeronautics and Astronautics, pp.1-9, 2011.
[10] M. Nagai and M. Shino, “Yaw-Moment Control of Electric Vehicle for Improving Handling and Stability”, *JSAE Review*, Vol. 22, No. 4, pp.473-480, 2001.
[11] M. Abe, Y. Furukawa, Y. Kano, K. Suzuki and Y. Shibahata, “Side-slip Control to Stabilize Vehicle Lateral Motion by Direct Yaw Moment”, *JSAE Review*, 22, No. 4, pp.413-419, 2001.
[12] W.-Y. Hsu, *Design of a Robust Self-Tuning Fuzzy PID Controller*, Taiwan Kaohsiung, National Kaohsiung University, 2010.
[13] K. H. Ang, G. C. Y. Chong and Y. Li, “PID Control System Analysis, Design, and Technology”, *IEEE Trans. Control Syst. Technol.*, Vol. 13, No. 4, pp.559-576.

Searching Fuzzy Information in Digital Library

Do Quang Vinh
Deparment of Information Technology
Hanoi University of Culture
418 La Thanh Street, Dong Da, Ha Noi, Viet Nam
Email: vinhdq@huc.edu.vn

ABSTRACT

Now, method of retrieving and collecting information has changed. It's not necessary to go out for searching and accessing large available information online via portal, provided by several information providers such as: DL, digital publishers, enterprises, organizations, individuals. Accessing information is not limited by available books or magazines in the nearest library, it can be accessed via big databases and documents distributed in over the world. It's not only texts and digital data, but also includes images, sounds/voices, geographic data, video, audio, multimedia. It enables users to go a virtual travel in the museums, historic spots and wonder of natures, to participate in the concerts and virtual performance, to watch movies and read books, to listen the lectures and music - all via DL.

KEYWORDS

digital library, model of retrieving information, method of retrieving information, retrieving fuzzy information, searching with near operator

1. DEFINITION OF DIGITAL LIBRARY

1.1 Informal Definition

Herein, we introduce informal definitions on DL.

Definition 1 (Arms W.Y.) [1]: digital library is a managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network. The main idea is the managed information. Digital libraries contain diverse collections of information for use by many different users. Digital libraries range in size from tiny to huge. They can use any type of computing equipment and any suitable software. The unifying theme is that information is organized on computers and available over a network, with procedures to select the document in the collections, to organize it, to make it available to users, and to archive it.

Definition 2 (Reddy R., Wladawsky-Berger I.) [3,10]: DL is network document storage on digital documents, image, sound, science data and software that are the core of today's Internet and tomorrow's universally accessible digital repositories of all human knowledge.

Definition 3 (The Digital Library Federation) [3,10]: Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities.

In conclusion, Digital Library is a huge managed collection of digital information with associated services.

1.2 Formal Definition

Next, we introduce formal definition on DL:

Digital Library is a set of four (R, MC, DV, XH) in which:

- R is a storage;
- MC is index of meta-data;
- DV is a collection of services containing indexing, searching and browsing services;
- XH is a user community of digital library.

2. MODEL OF INFORMATION RETRIEVAL

Information retrieval (IR) refers to managing, storing, searching and evaluating relevant information to user's demand. [2, 4, 5, 6, 7]

Overall model of information retrieval is a pair including objects and an search mapping to some objects with an representative object for a query.

Given

$$D = \{d_1, d_2, ..., d_M\}, M \geq 2 \quad (1)$$

is a limited non-empty collection of object.

Note: case of M = 1 can be considered, but it's normal. Typical objects are representation.

Given $\Re$ as a search mapping from D in $\rho(D)$, it means that,

$$\Re : D \rightarrow \rho(D). \quad (2)$$

By combining collection of D objects and $\Re$ search mapping, we define structure of information retrieval as follows:

Definition 3.1 (structure of information retrieval):

Structure of information retrieval (SIR) is a set 2 $S = <D, \Re>$. (3)

Definition 3.1 is a general definition: it doesn't mention about distinct type of $\Re$ search mapping and object D. So, it can be obtained the different types of model IR by specifying D and $\Re$.

We introduce a consistent definition on IR models using SIR.

Definition 3.2 (Model of information retrieval - MIR):

Model of information retrieval (MIR) is a SIR $S = <D, \Re>$ with 2 following attributes:

(i) $q = \delta \Rightarrow \mu\tilde{a}_i(q, \delta) = 1 \ \forall i, q, \delta$ (mapping); (4)

(ii) $\Re^i(q) = \{\delta \in D | \mu\tilde{a}_i(q, \delta) = \max \mu\tilde{a}_k(q, \delta_k)\} \cap a\alpha_i$, optional fixed i.

in which:

+ $T = \{t_1, t_2, ..., t_N\}$ is a limited collection of index term, $N \geq 1$;

+ $O = \{o_1, o_2, ..., o_U\}$ is a limited collection of object, $U \geq 2$;

+ $(D_j)_{j \in J = \{1, 2, ..., M\}}$ is a cluster group of object, $D_j \in \rho(O)$, $M \geq 2$;

+ $D = \{\delta_j | \ j \in J\}$ is a collection of document, in which fuzzy collection is standardized $\delta_j = \{(t_k , \mu\delta_j(t_k)) | \ t_k \in T, k = 1, ..., N\}$, $j = 1, ..., M$, $\mu\delta_j : T \rightarrow S \subseteq [0, 1] \subset \mathbf{R}$ is cluster representation of cluster object D_j . For example, O may include articles, each article is a cluster and each of fuzzy representation of cluster is a document. In this case, if fuzzy collection is exact collection, the document is unique to presentation of classic binary vector. Other example, cluster can be a collection of related articles, in which representation of cluster or document is a fuzzy representation for one of articles or false articles (or it was not an article in cluster but it has good description on the whole content of cluster). Therefore, it is a special case of traditional model of IR cluster.

+ $A = \{\tilde{a}_1, ..., \tilde{a}_C\}$ is a limited standard collection, $C \geq 1$, in which $\tilde{a}_i = \{((q, \delta_j), \mu\tilde{a}_i(q, \delta_j)) | \ \delta_j \in D, j = 1, ..., M\}$, $i = 1, ..., C$ is a fuzzy standardized relation, $\mu\tilde{a}_i : D \times D \rightarrow [0, 1] \subset \mathbf{R}$, $q \in D$ optional fixed. In addition, classic IR has splitting attribute (bipolar) in which there are 2 clear standards:

(i) existence and non-existence;

(ii) searching depends on (i).

We assume that it has more than 2 standards (or relevant, irrelevant, undeterminable) with different levels. So, we have to accept the standard of fuzzy relation.

+ $a\alpha_i = \{\delta \in D | \ \mu\tilde{a}_i(q, \delta) > \alpha_i\}$, $i = 1, ...,$ C is a α_i-strong standard section $\tilde{a}_i$, $\alpha_i \geq 0$, $q \in D$ optional fixed;

+ $\Re : D \rightarrow \rho(D)$ is a search mapping. In general view, searching is link of subset document and a query if they are linked - according to a strong enough selection standard. So that, we have to consider that query which is a document and search, is defined to use α-section.

Next, we will introduce definition of model of information retrieval of R.B. Yates and B.R. Neto [10]:

Definition 3.3:

A model of information retrieval is set of four $[D, Q, F, R(q_i, d_j)]$

in which:

+ D is a collection of document;

+ Q is a collection of user query;

+ F is a simulation framework of document, query and relation presentation;

+ $R(q_i, d_j)$ is a sorting function to link a real number with a query $q_i \in Q$ and a document representation $d_j \in D$. The sorting function determines order between documents and query q_i.

3. FUZZY INFORMATION SEARCH

3.1 Definition

Fuzzy search is searching information with incomplete input together with user's expectation or getting relative result with desired data. The reason we need studying and developing the fuzzy search:

+ User don't remember exactly searching term

+ Digital documents contain spelling errors when editing

3.2 Searching with Near Operator

In general, we search a character string by inputting exactly that string. For example, when we want to search information of Bill Gates, our character string query is “William Henry Gates”. However, search engine will not give out pages which have the same information on Bill Gates, only include string “William H. Gates” or “William Gates”. In order to solve this matter, we can use OR operator as follows: “William Henry Gates” OR “William H. Gates” OR “William Gates”.

To replace using many OR operators as above and improve searching efficiency of desired Web, some search engines have developed NEAR operator. Herein, NEAR operator means searching pages contain nearby words. For example, “William NEAR Gates”. How exactly "near" will depend on each specific search engine in the digital library.

For example: query “Digital Near/3 Library” in the IEEE digital library, search engine queries Web pages containing Digital and Library words not exceed 3 words. [7, 9]

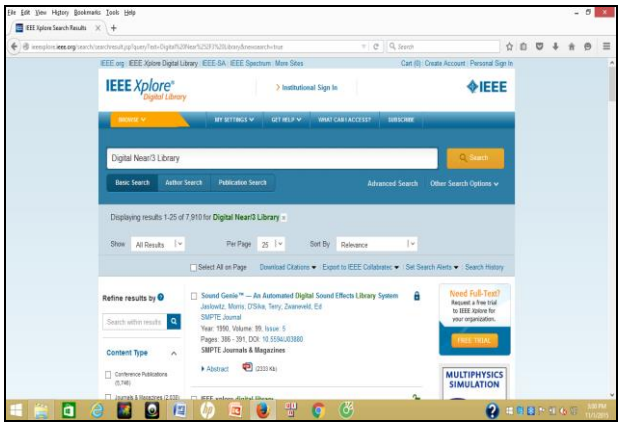


Fig. 1 – Operator NEAR in the IEEE digital library

3.3 Searching by Root Words

Progress of searching by root words is root-finding allows to spend less time on comparing query terms to index terms, for example both "computation" and "computer" are accepted to equal with "compute". Root-finding progress which eliminates one or more suffix(es) from word to reduce into root words, converts to neutral term without tense and plural state. The easiest way to evaluate exactly root-finding progress is considering some sample texts. In order to create index terms, all punctuations should be eliminated and as above discussion, all letters are grouped into word form.

At a glance, result of root-finding progress is not much value - the algorithm is not value. However, the first note, the last expression of root words is not important if it's unique to class of root terms and the second note, the alternation is repeated. Of course, it's not necessary to reverse the alternation because we only use root-finding progress to create index terms and texts are also stored exactly; the root word doesn't need to be any meaningful English word. In fact, "computer", "computing" and "computation" can be changed into "ppzxg" by root-finding progress if there is not any other word.

Root-finding progress is not necessary to fit with all elements of database. If there is a book in the list of library and putting the exact query "Arms AND Digital AND Libraries", root-finding progress will be mapped into one phrase with "Arms AND Digital AND

ISBN: 978-1-941968-42-0 ©2017 SDIWC

Library" and it may result in polysemy. In author field of directory database, it's suitable to disable root-finding progress; limit the root-finding progress with certain section of document which is optional, is available with database designer.

Actual installation of root-finding progress requires deep knowledge on language. For example, for English, in some cases: suffixes such as: -s, -ed, -ing, -ly and etc. are easy. In the other hand, exception of simple rule was recognized and supplemented by deep rules to prevent them being searched by root word; after that, they recognized other exceptions one by one etc. Root-finding progress contains more than 500 rules and exceptions, encoding as a limited state.

The main reason of root-finding and word-grouping progress is to simplify query formation. However, it causes significantly size of IF (Inverted File) becoming an advantage. There are 2 reasons: firstly, it has less stored IL (Inverted List) and they becoming denser, and storing price bases on one pointer is cheaper; secondly, average frequency of term inside the document tends to increase, it means that there is less pointer.

For example, after processing word-grouping and root-finding for TREC database, the number of pointers decrease about 16%, distinct terms decrease about 40% and total space which is used by Golomb encoding, decreases about 30%. However, the saving by using index of root-finding which is compensated up to a certain level equaling value of storing vocabulary by root-finding progress. A vocabulary which is not processed by word-grouping or root-finding, can be shared among compact element of document and sub-index. It can not be done by a root-finding progress. In TREC, vocabulary requires about 5 MB, so it's suitable for processing root-finding with amount of saving about 35 MB.

In a SF (Signatured File), effect of processing root-finding and word-grouping decrease the number of distinctive terms appearing in each record, so it's necessary to reduce width of signature for a given matching, hence accumulating the saving.

3.4 Searching Synonym Words

Any natural language has synonym words. When searching information on certain matter, we input typical keywords and get results from search engine. However, not only web pages contain query keywords including desired content, but also other web pages including these contents. In general, these web pages contain synonym words with query keywords. Hence, they studied and developed searching with synonym words of search engine.

Thus, searching with synonym words of search engine building set of synonym words dictionary according to supported language. When searching, search engine will find all web pages containing synonym words with query keyword. For example, when searching information on web crawler, search engine will find all web pages containing information on web robot, spider etc.

The advantage of searching synonym words is that users don't need to remember all synonym words. Especial in new technology fields, there are many terms used for a same matter that users can not know about it.

4. CONCLUSION

Nowadays, DL becomes important in national and international aspects because of information explosion follow by exponential function on Web. Web interface develop from browsing to searching, everybody do searching on Web everyday in the world. Thus, they have been concentrating on studying and developing technologies of searching in big database. There are many databases distributed all over the world where each of small group maintains document database by oneself. In DL, not only searching text information, but also searching multimedia information, search engines have been studied and improved, in which including method of retrieving fuzzy information.

APPENDIX: [8]

Algorithm of Searching Fuzzy Information

```
Search (T,n,p,m,k)
{
        /* Processing */
For each c ∈ Σ
{
T[c] = (c != p[m]) (c != p[m+1])
… (c != p[1]);
T[c] = 0s_{k+1}(t[c], 0) 0s_{k+1}(t[c],1)
…
        0s_{k+1}(t[c], m-k-1)
S[c] = ( c ∈ p[1..k+1]);
}
D_in = (0 1^{k+1})^{m-k} ;
M1 = (0^{k+1} 1)^{m-k} ;
M2 = (0^{k+1} 1)^{m-k-1} 0 1 1^{k+1} ;
M3 = 0 ^{(k+1} 0 1^{k+1} ;
G = 1 << k ;
/* Searching */
D = D_in ;
i = 0 ;
 while ( ++i <= n)
        if  (S[T[i]]) do
        {x  =  (D  >>  (k+2))  |
S[T[i])
        D = (( D << 1) | M1 ) &
((   D   <<   (k+3))   |   M2   )
        & (( x + M1 ) Λ x ) >> 1
) & D_in
        if  ( D & G == 0)
        {
                D = D | M3
        }
}
while    ( D != D_in && ++i <=n )
```

REFERENCES

[1] Arms W.Y. (2003), Digital Libraries, MIT Press, Cambridge.

[2] Chowdhury G.G. (1999), Introduction to Modern Information Retrieval, Library Assocciation Publishing, London.

[3] Lesk M. (2005), Understanding Digital Libraries, 2nd Edition, Morgan Kaufmann, San Francisco.

[4] Large A., Tedd L.A., Hartley R.J. (2001), Information Seeking in the Online Age, K.G. Saur Verlag, Munchen.

[5] Korfhage R.A. (1997), Information Storage and Retrieval, John Wiley, New York.

[6] Kowalski G. (1997), Information Retrieval Systems, Kluwer Academic Publishers, Boston.

[7] Schatz B.R., "Information Retrieval in Digital Libraries", Science 275, 1997, pp. 327-334.

[8] Wiederhold G. (2001), Database Design, 2nd Edition, McGraw-Hill, New York.

[9] Nguyễn Như Phong (2005), Lý thuyết mờ và ứng dụng, Nxb Khoa học và Kỹ thuật, TP. Hồ Chí Minh.

[10] Đỗ Quang Vinh (2009), Thư viện số: chỉ mục và tìm kiếm, Nxb Đại học Quốc gia Hà Nội.

Automated Processing for Color Image Arrangement Based on Histogram Matching Using Gaussian Distribution

Yusuke Kawakami
DynaxT Co., Ltd.
2217-6 Hayashi, Takamatsu City, Kagawa 761-0301, Japan
riverjp2002@gmail.com

Tetsuo Hattori, Yoshiro Imai, Kazuaki Ando, Yo Horikawa,
Graduate School of Kagawa University
2217-20 Hayashi, Takamatsu City, Kagawa 761-0396, Japan
hattori@pe.kagawa-u.ac.jp, {imai, ando, horikawa}@eng.kagawa-u.ac.jp

R. P. C. Janaka Rajapakse
Tainan National University of the Arts
66 Daci, Guantian District, Tainan 72045, Taiwan
janakaraja@gmail.com

ABSTRACT

This paper describes a method for color image arrangement using Histogram Matching based on a parameter estimated Gaussian distribution (HMGD). In our previous papers, we have presented some methods for the estimation of parameters in Gaussian distribution from the curvature computation of the original cumulative histogram. However, since the histogram of original image is not always ideally shaped for the Gaussian distribution approximation, the parameters estimation method based on curvature computation may not work well. In this paper, we propose an improved parameter (especially variance) estimation method using regression analysis. This method firstly detects the part of peak in original image histogram by using curvature computation, and secondly carries out the regression analysis based on approximated curvature formula. This paper also shows the experimental results.

KEYWORDS

Image processing, Curvature, Variance estimation, Histogram matching, HMGD

1 INTRODUCTION

Recently, automated color image processing for arrangement or enhancement has been very popular such as in Digital Signage, Smart Phone, etc [1-3].

In our previous papers, we presented that the Histogram Matching processing based on Gaussian distribution (HMGD) is a kind of automated image arrangement method using Elastic Transformation [4-5] on the brightness axis. And we illustrated from the comparative investigation, that HMGD processing has high possibility to give better feeling impression than that of original image [6]. Then, aiming to improve the HMGD processing, we have pursued the method to precisely estimate the parameters (i.e., average and variance) in Gaussian distribution by curvature computation.

However, since the histogram shape of original image is not always ideal for the Gaussian distribution approximation, the variance estimation using the curvature computation may not work well.

In this paper, we propose an improved parameter (especially variance) estimation method using regression analysis. This method firstly detects the part of peak in original image histogram by using curvature computation, and secondly performs the regression analysis based on approximated curvature formula. This paper also shows the experimental results.

2 PRINCIPLES

2.1 Brightness Peak Detection of Original Image

In the section, we describe the principle of brightness peak detection of original image.
The Histogram Matching based on the Gaussian distribution [4-9] processing need to calculate transforms function for brightness peak of histogram. And the solution to detect it is curvature computation of the histogram.
Let y be a function with respect to x, the definition curvature $R(x)$ is given by Eq.(1) [6-9].

$$R(x)=\frac{d^2y}{dx^2}\Bigg/\left(1+\left(\frac{dy}{dx}\right)^2\right)^{\frac{3}{2}}. \tag{1}$$

And let $g(x)$ and K be Gaussian distribution function and a coefficient which is defined by following equation.

$$\begin{aligned} &g(x)=\frac{K}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right), \\ &\frac{K}{\sqrt{2\pi\sigma^2}}\int_0^L\exp\left(-\frac{(u-a)^2}{2\sigma^2}\right)du=1. \end{aligned} \tag{2}$$

Next, let y (=$f(x)$) be a function representing the cumulative histogram which is represented Eq. (3). That is, dy/dx and d^2y/dx^2 be described as Eq. (4) and (5), respectively. From Eq. (4) and (5), we obtain the approximation of curvature $R(x)$ as Eq. (6).

$$f(x)=\int_0^x g(u)du=\frac{K}{\sqrt{2\pi\sigma^2}}\int_0^x\exp\left(-\frac{(u-a)^2}{2\sigma^2}\right)du. \tag{3}$$

$$\frac{dy}{dx}=g(x)\frac{K}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(u-a)^2}{2\sigma^2}\right). \tag{4}$$

$$\frac{d^2y}{dx^2}=\frac{dg(x)}{dx}=\frac{(a-x)}{\sigma^2}g(x) \tag{5}$$

$$R(x)=\frac{\frac{(a-x)}{\sigma^2}g(x)}{\left(1+g(x)^2\right)^{\frac{3}{2}}}\approx\frac{(a-x)}{\sigma^2}g(x). \tag{6}$$

From Eq.(6), we understand that the curvature $R(x)$ varies the sign according to the value of x [9]. That is, if $x<a \rightarrow R>0$ (downward convex shape), and if $x>a \rightarrow R<0$ (upward convex shape).

2.2 Variance Estimation Based on Regression Analysis

In this section, we describe how to optimize the shape of the reference histogram, which is used in the HMGD processing [9]. We explain the proposed method which is based on the regression analysis.
Figure 1 shows the conceptual image of the original image histogram which is variance σ^2 and average a. And Figure 2 shows its cumulative histogram.
From Eq. (6), we can describe $R(x)$ as Eq.(7).

$$R(x)\approx\frac{(a-x)}{\sigma^2}g(x)=\frac{1}{\sigma^2}(a-x)g(x). \tag{7}$$

Then, let $C = 1/\sigma^2$ and $H(x) = (a\text{-}x)g(x)$ respectively, we can derive Eq. (8).

$$R(x)\approx CH(x) \tag{8}$$

Now, we can calculate a constant C by using least square regression analysis method [10] following Eq.(9) through (10).

$$\begin{aligned} &R_i=CH_i+\varepsilon_i, \\ &\varepsilon_i\sim N(0,\sigma^2)\quad(i=1,\ldots n) \end{aligned} \tag{9}$$

$$C=\frac{\sum_{i=1}^{n}(H_iR_i)}{\sum_{i=1}^{n}(H_i)^2},\quad \sigma^2=\frac{1}{C} \tag{10}$$

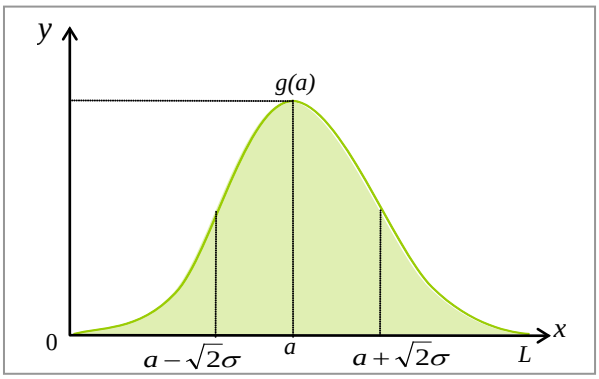


Figure 1. Conceptual image of the histogram [8-9].

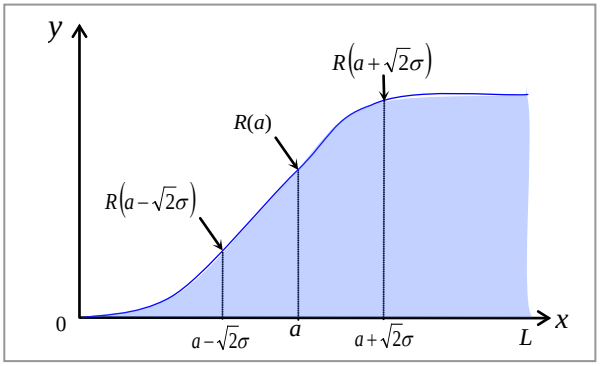


Figure 2. Conceptual image of the cumulative histogram [8-9].

3 Experimental Results

Figure 3 shows the original image (grey level image and color one) and the corresponding histogram and its cumulative one. Similarly, Figure 4 shows the HMGD processed images (grey level image and color one) by variance estimation using regression analysis and the corresponding histogram and its cumulative, where we performed regression analysis increasing the number of sample points (i.e., *i*) by 4. From Figure 4, we consider that, if the σ of the Gaussian distribution of reference histogram is increased, the HMGD processing result goes better and becomes natural. However, if we take a lot of sample points, there seems to be a tendency that the tone of HMGD becomes vivid but a little bit unnatural comparing to the original image.

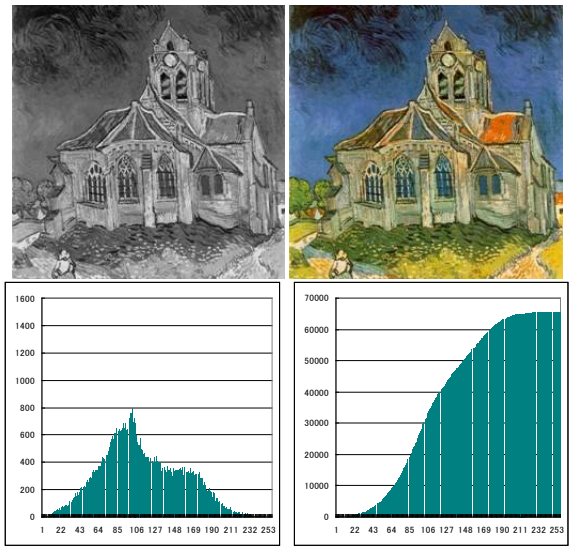
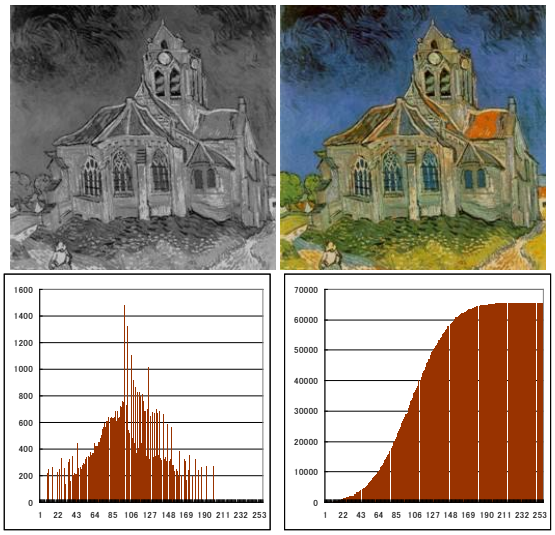
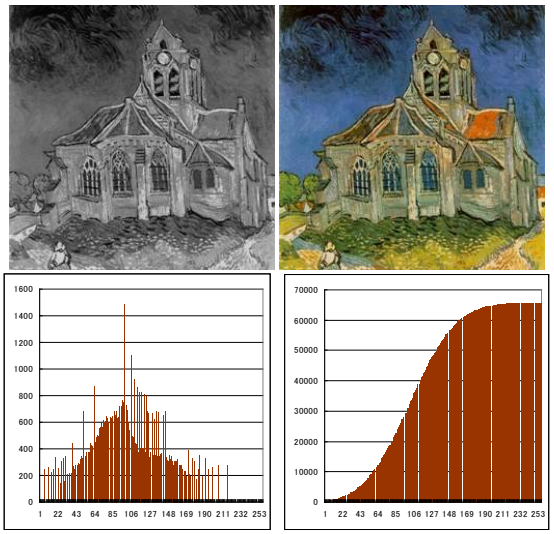


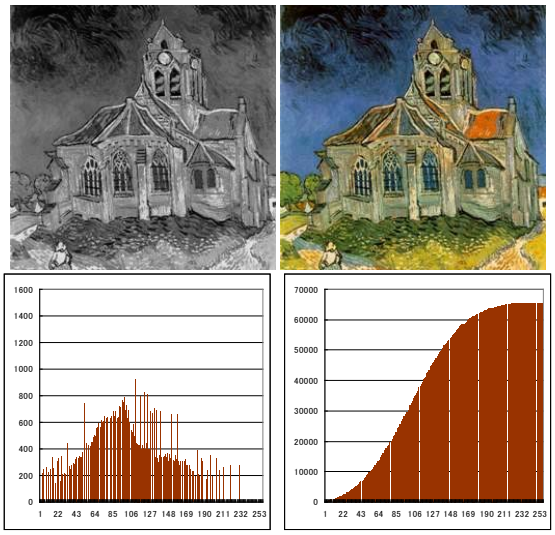
Figure 3. Original images (grey level and color) and the corresponding histograms of the greay level image (histogram and cumulative).



(a) HMGD processing which applied regression analysis based variance estimation (*i*=16, σ=37.796)



(b) HMGD processing which applied regression analysis based variance estimation (*i*=20, σ=43.596)



(c) HMGD processing which applied regression analysis based variance estimation (*i*=24, σ=49.797)

Figure 4. Example of results and the correspondence histograms (2 of 2).

ISBN: 978-1-941968-42-0 ©2017 SDIWC

4 Conclusion

In this paper, we have proposed the method to estimate the parameter (especially variance) of Gaussian distribution making regression analysis from the original image histogram, in order to use the reference histogram in HMGD processing.

This method firstly detects the part of peak in original image histogram by using curvature computation, and secondly performs the regression analysis based on approximated curvature formula.

From the experimental results, we understand that the variance estimation based on the regression analysis depends on the number of sample points.

For example, if we take too less sample points, HMGD processing gives reduced contrast and then color tone becomes unnatural On the other hand, if we take too much sample points, it seems that the results tend to be vivid and unnatural comparing to original one.

As a further study, we have to obtain the decision making method as to how many sample points will be appropriate for the regression analysis based variance estimation.

REFERENCES

[1] R. C. Gonzalez and R. E. Woods, Digital Image Processing, Addison-Wesley Publishing Company, 1993.

[2] B. Jahne, Digital Image Processing --Concepts, Algorithms, and Scientific Applications-- 4th edition, Springer, 1995.

[3] E. S. Umbaugh, Computer Vision and Image Processing: A Practical Approach Using CVIP tools, Prentice Hall PTR, 1998.

[4] W. Burger and J. M. Burge, Principles of Digital Image Processing: Fundamental Techniques, Springer 2009.

[5] T. Izumi, T. Hattori, S. Sugimoto, and T. Takashima, "Color Image Arrangement Using Elastic Transform on Principal Component Axis," Journal of Japan Society of Kansei Engineering, Vol. 8, No. 3, pp. 667-674, 2009. (in Japanese)

[6] Y. Kawakami, T. Hattori, D. Kutsuna, H. Matsushita, Y. Imai, H. Kawano, R.P.C. Janaka Rajapakse, "Automated Color Image Arrangement Method Based on Histogram Matching - Investigation of Kansei impression between HE and HMGD -," International Journal of Affective Engineering, Vol. 14, No. 2, ISSN 2187-5413, pp. 85-93, 2015.

[7] Y. Kawakami, T. Hattori, Y. Imai, H. Matsushita, H. Kawano, R. P. C. Janaka Rajapakse, "Kansei Impression and Automated Color Image Arrangement Method," Journal of Robotics, Networking and Artificial Life, Vol. 1, No. 1, ISSN 2352-6386, pp. 60-67, 2014.

[8] Y. Kawakami, T. Hattori, Y. Imai, H. Matsushita, H. Kawano, and R. P. C. Janaka Rajapakse, "Automated Color Image Arrangement Method Using Curvature Computation in Histogram Matching," Proceedings of International Conference on Artificial Life and Robotics (ICAROB 2015), ISBN 978-4-9902880-9-9, pp. 272-277, Oita, Japan, 2015.

[9] Y. Kawakami, T. Hattori, Y. Imai, Y. Horikawa, H. Matsushita, R. P. C. Janaka Rajapakse, "Automated Processing of Multiple-Brightness Peak Histogram Image Using Curvature and Variance Estimation," Journal of Robotics, Networking and Artificial Life Vol. 3, No. 1, ISSN 2352-6386, pp. 55-60, 2016.

[10] D. G. Kleinbaum, L. L. Kupper, A. Nizam, E. S. Rosenberg, Applied Regression Analysis and Other Multivariable Methods, Brooks/Cole Pub Co, 2013.

An Algebraic Approach for the Detection of Vulnerabilities in Software Systems

Oleksandr Letychevskyi and Vadim Sukhomlinov
Glushkov Institute of Cybernetics of
National Academy of Sciences (Ukraine),
Intel Corporation (USA)
lit@iss.org.ua
vadim.sukhomlinov@intel.com

ABSTRACT

The paper presents an algebraic approach for finding vulnerabilities in a program system that is given as the sequence of processor instructions. The main result of the paper is the transformation of code to algebraic specifications and providing its symbolic modeling for the detection of vulnerability cases that are presented as formulas in logic language. The method anticipates the usage of solving and proving systems integrated with the Algebraic Programming System developed by the authors. A given example illustrates the method.

KEYWORDS

Cybersecurity, algebraic programming, predicate transformer, vulnerability, safety, symbolic modeling.

1 INTRODUCTION

Detection of vulnerability and software security analysis is now one of the most challenging problems in software engineering. A great amount of tools have been developed since the 1990's to meet the requirements for prevention of intrusion and security violations. One of the most popular methods of vulnerability detection is a static method that analyzes the code and offers a set of possible points where a problem could occur. A good survey of static tools is presented in [1]. The static method has been realized in a number of tools, such as Parasoft C++ [2], Klocwork [3], and is available today in mainstream development tools such as Clang (LLVM) [4] and Microsoft Visual Studio [5]. It can detect various types of security vulnerabilities in programs like buffer overflow and null pointer assignment, but many other types of vulnerabilities are very difficult to find. Moreover, the static method can list a big variety of false issues and can fail to detect the issues that are not presented in the given code, such as library functions.

One of the ways to find all vulnerabilities is to run an exhaustible execution of all program scenarios by taking into account all input values. For this purpose, the symbolic execution approach was used in analysis of a Java program [6]. Such an approach demands the usage of powerful solving and proving machines that are actively used in a model-checking domain. Nevertheless, symbolic execution of the program on the level of the Java language still does not resolve the vulnerability detection problem due to the impossibility for checking compiled components like libraries or third-party tools.

Fuzzing, another automated method for software testing, has roots from the 1950's, when data were still stored on punched cards. Programmers used punched cards that were pulled from the trash or card decks of random numbers as input to computer programs [7]. If an execution revealed undesired behavior, a bug had been detected and was fixed. In the urge to enhance security, fuzzing is commonly used today in automated building/integration solutions. Compilers provide support in instrumenting code to detect leaks, out-of-bound accesses, etc., such as AddressSanitizer in Clang. Special fuzzers such as AFL and libFuzzer [8] feed inputs.

Academic research focused on improving coverage and automating crash analysis.
We propose to resolve this problem by the symbolic execution of the instructions of a program executable code by using special slicing methods that are popular in model-checking techniques. Symbolic execution framework is based on the Algebraic Programming System (APS) [9] that has been developed and maintained since the 1990's at the Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine.
In this paper, we will adhere to the following content. First, we will formulate the problem statement and give the high level scheme of the method. We then will present the main algebraic notions that will be necessary for the understanding of the method and the technology chain to be used. Finally, we will illustrate the algorithm by the example of vulnerability detection presented in the Common Vulnerabilities and Exposures (CVE) database [10].

2 PROBLEM STATEMENT AND HIGH LEVEL DESIGN OF METHOD

Our goal is to prove that the known vulnerability can be reached or not in a given system. We consider as "known" the vulnerabilities from CVE. The formalized case of vulnerability, together with a model of executable code are the input of the technology.
It is anticipated that the case of vulnerability will be formalized manually as a formula in some formal language. The examples of vulnerability formalization will be considered further, and the language for presentation of the formula should express the static properties (logic language), behavioral properties (process algebra) or time properties (temporal logic).
The executable code itself also is presented by means of formal specifications that are the model of the program. These specifications are the inputs for technology and corresponding transformation that should be provided.

Here, shown below, is the scheme that illustrates the process of detection of vulnerability. It consists of two stages.

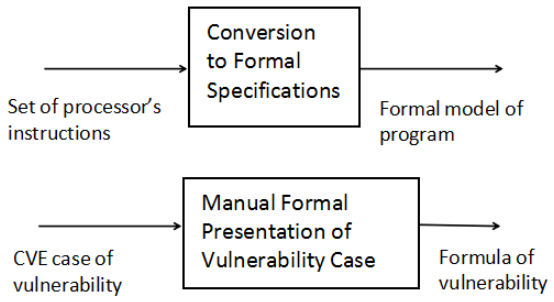


Figure 1. The first stage of method. Input data preparation.

During the first stage we should disassemble the input code that is uploaded to the memory. Reading the instructions directly from memory, we can process the part of the system that contains third-party tools or libraries as is. In the development environment, assembler code can be directly produced by the compiler.
After the preparation, we can use the Algebraic Programming System and its component for further processing in stage 2.

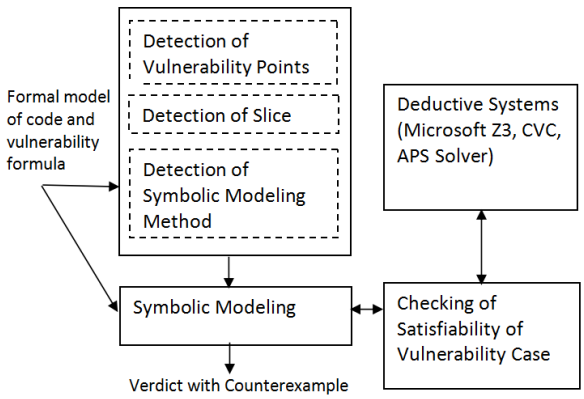


Figure 2. The second stage of method. Symbolic execution and detection of presence of vulnerability.

Before direct symbolic modeling we should detect the possible points of intrusion for security of data. Usually these are the points of input in the program system. The explored property of vulnerability also defines the subset of processor instructions to be

processed, so we can symbolically execute only the interested part of program by selecting the method of symbolic modeling.

Symbolic modeling of the selected program slice is performed in the scope of the APS system, that uses as the third-party solvers as the integrated ones.

3 ALGEBRAIC MODEL OF EXECUTABLE CODE

3.1 Behavior Algebra

Algebra of behavior was developed by D. Gilbert and A. Letichevsky in 1997 [11]. It was realized in the scope of the Insertion Modeling System (IMS) as an extension of APS. Behavior algebra is a two-sorted universal algebra. The main sort is a set of behaviors and the second sort is a set of actions. The algebra has two operations, three terminal constants, and a relation of approximation. The operations are the prefixing *a.u* (where *a* is an action, and *u* is a behavior) and non-deterministic choice of behaviors *u + v* (associative, commutative, and idempotent operations on the set of behaviors). The terminal constants are successful termination Δ, deadlock *0*, and non-determinate behavior $\perp$. The relation of approximation $\sqsubseteq$ is a partial order on the set of behaviors with minimal element $\perp$. The following example of behavior expressions

B0 = a1.a2.B1 + a3.B2,
B1 = a4.$\perp$,
B2 =…

means that behavior *B0* could be interpreted as sequence of action *a1*, *a2* and the behavior *B1* afterwards, or the action *a3* with the next behavior *B2*. Behavior *B1* will finish after action *a4*.

3.2 Basic Protocols Language

The Basic Protocols language has been developed in the scope of the Verification of Requirements Specifications (VRS) project [12], implemented together with Motorola and the Glushkov Institute of Cybernetics.
The language is built over some attribute environment, where agents interact one with another. Every agent is defined by a set of attributes. An agent changes its state under some conditions formed by values of attributes. Every agent's actions defines some basic protocol that is a triple: $B = \langle P, A, S \rangle$, where *P* is a precondition of basic protocol presented as a formula in some basic logic language, *S* is a postcondition, and *A* is a process that illustrates agent transition. As a basic logical language, we consider the set of formulas of first-order logic over linear arithmetic. As a whole, the semantic of a basic protocol means that the agent could change its state if the precondition is true and the state will change correspondingly to the postcondition, which is also a formula of first-order logic. The postcondition could also contain an assignment statement.
The process of basic protocol depends on the subject domain and illustrates the sequence of basic protocols application. In a telecommunications domain, it could be the sending or receiving of signals with corresponding parameters. In the current case it is given as an identifier of a basic protocol.

3.3 Semantic of Processor Instructions by Means of Behavior Algebra and BP-Language

The description of the architecture of the Intel 64 and IA-32 processors is presented in [13]. We consider the interactions of the following agents: processor, memory, and external environment. The architecture is composed of the attribute environment, where attributes are the set of general purpose registers (AH, AL, AX, EAX, RAX,…) of different types (byte, word, doubleword,…), and different bit capacities. Moreover, we consider as attributes the set of flags that are contained in the EFLAGS/RFLAGS register.
In a huge amount of instructions we distinguish:
- control flow instructions (JCC, JMP, CALL, …) that provide navigation via

program code corresponding to attributes values;
- instructions that change the attributes environment. This is a set of ALU instructions. These instructions change the values of registers or memory, can provide calculation, and compare values in registers with settings of corresponding flags.
We transform the sequence of instructions into behavior algebra expressions with actions that are the basic protocols with preconditions containing predicates and postconditions that define changing attributes. For example, branch instruction:

```
40a984: jne 40ab50
```

could be converted to a behavior algebra expression, covering possible outcomes based on the state of the *ZF* flag in EFLAGS:

B40a984=a_jne1.B40ab50+a_jne2.B40a98a

where the actions are the following:

a_jne1 = (ZF=0) → <“jne1”> 1
a_jne2 = ~(ZF=0) → <“jne2”> 1

We denote behavior identifiers together with the hexadecimal address of instructions in a program segment for traceability to assembler code. The expression above means that the instruction by the address `40a984` will pass the control to the instruction by the address `40ab50` if flag *ZF* is equal to *0*; otherwise the next instruction by the address `40a98a` will be performed. The condition of transition is in a predicate that is the precondition of the basic protocol. The postcondition is absent so it is equal *1*. The process of basic protocol is given as some string with the name of an action containing the instruction. It will be used for traces defining the program behavior. The instructions that change the attributes could be presented also as basic protocols with postcondition containing this changing. For example, the instruction:

```
40a99d: add DWORD PTR [r13 +
0x15c],r8d
```

will be transformed to

B40a99d = a_add_3480.B40a9a4

where

a_add_3480 = 1→ <”a_add_3480”>
Memory(r13 + 348) := Memory(r13 + 348) + r8d;
ZF := (Memory(r13 + 348) + r8d = 0);
SF := (Memory(r13 + 348) + r8d < 0);

This instruction performs the adding of a memory element that is available by the given address in register r13 to the content of doubleword register r8d. The given flags will be set to bit 1 or 0 corresponding to the truth of the given equality or inequality. There are other flags (e.g., CF, OF, AF, PF, etc.) that are affected, but these are not illustrated for simplicity.
All this semantic of instructions has been defined directly following specification in the data sheet, and we see that formalization of executable code is not that complicated for representation by formal logic language. Consider the following code fragment:

```
0000000000425060 <SSL_CTX_use_certificate_file>:
  425060:  41 55                 push   r13
  425062:  41 54                 push   r12
  425064:  49 89 f5              mov    r13,rsi
  425067:  55                    push   rbp
  425068:  53                    push   rbx
  425069:  49 89 fc              mov    r12,rdi
  42506c:  89 d5                 mov    ebp,edx
  42506e:  48 83 ec 08           sub    rsp,0x8
  425072:  e8 d9 24 fe ff        call   407550 <BIO_s_file@plt>
  425077:  48 89 c7              mov    rdi,rax
  42507a:  e8 a1 31 fe ff        call   408220 <BIO_new@plt>
  42507f:  48 85 c0              test   rax,rax
  425082:  0f 84 b0 00 00 00     je     425138
<SSL_CTX_use_certificate_file+0xd8>
  425088:  4c 89 e9              mov    rcx,r13
```

Figure 3. Example of code

which could be translated to algebra behavior expressions

```
B425060 = a_push_33766.B425062,
B425062 = a_push_33767.B425064,
B425064 = a_mov_33768.B425067,
B425067 = a_push_33769.B425068,
B425068 = a_push_33770.B425069,
B425069 = a_mov_33771.B42506c,
B42506c = a_mov_33772.B42506e,
B42506e = a_sub_33773.B425072,
B425072 = a_call_33774.call B407550.B425077,
B425077 = a_mov_33775.B42507a,
B42507a = a_call_33776.call B408220.B42507f,
B42507f = a_test_33777.B425082,
B425082 = a_je_33778.B425138 + a_alt_je_33779.B425088,
B425088 = a_mov_33780.B42508b,
```

Figure 4. Behavior expressions

The actions in behavior could be presented as the following.

```
a_push_33766 = Operator(1  -> ("x86: action 'push 425060';")
(rip := 4345954)),
a_push_33767 = Operator(1  -> ("x86: action 'push 425062';")
(rip := 4345956)),
a_mov_33768 = Operator(1  -> ("x86: action 'mov 425064';")
(rip := 4345959; r13 := rsi)),
a_push_33769 = Operator(1  -> ("x86: action 'push 425067';")
(rip := 4345960)),
a_push_33770 = Operator(1  -> ("x86: action 'push 425068';")
(rip := 4345961)),
a_mov_33771 = Operator(1  -> ("x86: action 'mov 425069';")
(rip := 4345964; r12 := rdi)),
a_mov_33772 = Operator(1  -> ("x86: action 'mov 42506c';")
(rip := 4345966; ebp := edx)),
a_sub_33773 = Operator(1  -> ("x86: action 'sub 42506e';")
(rip := 4345970; rsp := rsp - 8; ZF := (rsp - 8 = 0); PF :=
((rsp - 8) = 0); SF := (rsp - 8) < 0))),
a_call_33774 = Operator(1  -> ("x86: action 'call 425072';")
(rip := 4345975)),
a_mov_33775 = Operator(1  -> ("x86: action 'mov 425077';")
(rip := 4345978; rdi := rax)),
a_call_33776 = Operator(1  -> ("x86: action 'call 42507a';")
(rip := 4345983)),
a_test_33777 = Operator(1  -> ("x86: action 'test 42507f';")
(rip := 4345986)),
a_je_33778 = Operator((ZF = 1)  -> ("x86: action 'je
425082';")(rip := 4345992)),
a_alt_je_33779 = Operator(( ~(ZF = 1))  -> ("x86: action 'je
425082';")(rip := 4345992)),
a_mov_33780 = Operator(1  -> ("x86: action 'mov 425088';")
(rip := 4345995; rcx := r13)),
```

Figure 5. Actions of behaviors

Thereby the behavior expressions present the control flow of the program, and the actions define the changing of the attributes by means of the basic language. Such conversion gives the possibility of APS use where behavior algebra expressions are the input. In the scope of APS, we can provide symbolic modeling, analysis, generation of the scenario of program execution, and proving of the properties.

4 FORMAL PRESENTATION OF VULNERABILITY PROPERTY

Any property could be presented as some formula over attributes. The safety property states that something undesirable will never happen. In the context of vulnerability of a program, it could be reformulated that in some points of a program the truth of some formula over program attributes will never be achieved.

One of the easiest examples of the safety formula is the prohibition of the reference to not initialized memory, or "null pointer assignment." In the instruction's parameters `QWORD PTR[rax]`, the value in register rax shall not equal 0,

$$\sim(rax = 0)$$

When software does not validate input data properly, the attacker can craft the input with unexpected values, leading to undesirable behavior – arbitrary code execution, leakage of data or a crash. The condition for leakage of confidential data would be if the program allows memory access out of admissible bounds:

$$(Admissible0 \leq rax \leq AdmissibleN)$$

The famous "Heartbleed" vulnerability in OpenSSL library is an example of a lack of boundary check when a program reads adjacent memory. This example will be studied in detail in the next chapters.

Formalization of the safety property is the most complicated problem, and every case of vulnerability should be formalized separately. Some of the properties can be generated automatically, whereas some of them require manual formalization. This problem is open, but significant parts of known cases of vulnerabilities can be represented in formal logic.

5 SYMBOLIC MODELING

Given a model of executable code and the property that expresses the case of vulnerability, we use the symbolic methods for checking it that are implemented in APS.

The initial state of the program, especially values of registers and initial flags, can be presented by an initial formula. This formula can contain known (or predefined) values of attributes and unknown (arbitrary symbolic) values. Starting from the initial formula, we can apply the basic protocol corresponding to the control flow that is expressed in behavior algebra. The basic protocol is applicable if its precondition is satisfiable and consistent with the state of the environment. Starting from the formula of the initial state *S0* and from the initial behavior *B0*, we select the action and move to the next behavior. We check on the first step te satisfiability of the conjunction

$$S0 \wedge Precodnition(a1)$$

if *B0 = a1.B1*. The next state of the environment will be obtained by means of the predicate transformer; that is, the function over the state of the environment and the postcondition

PT(S0,Postcondition(a1)) = S1

The output is the new state of the environment expressed by the formula over the attributes.
Moving from the initial formula and applying the basic protocols, we will obtain the sequence of states or formulas over the attributes that define some possible trace of program execution. Using different traversal algorithms we can obtain the set of traces covering complete program behavior.
Searching for the intersection of the vulnerability formula with the satisfiable states of the program leads to a trace at an intersection, the results of which are potentially exploitable.

For a large system, this method might be unsuccessful due to the exponential explosion, so the suspected point might never be reached. To overcome this limitation, we can use backward symbolic modeling from the suspected point to the initial state. If all traces from the state presenting vulnerability lead to deadlocks, then vulnerability is unreachable.
If the APS dynamic method can be combined with static methods, the computation of invariants is possible in APS by different manners, especially the use of static detection of cycle invariants and methods of approximation of the invariant formula. In the case of approximation, we can compare the approximated formula with the vulnerability cases, and detect the issues earlier then the invariant will be computed. It should be taken into account that the problem of reachability is unresolved in a general way, so complete absence of vulnerability is not guaranteed.
Reduction states that the traversal could also be reached by use of those parts of the code that affect the formula of vulnerability. For this purpose, we can consider a slice (or subset) of behavior expressions and use only vulnerability formula attributes and its dependencies.

6 EXAMPLE

We consider the example of the “Open SSL Heartbleed vulnerability” that allowed an attacker to read sensitive data, such as authentication credentials, and secret keys by incorrect memory handling via Transport Layer Security (TLS) extension. The client requested the server to send strings from memory of a given length. The server sent the requested bytes, but the intruder could request lengths that were much greater than the buffer, so that the adjacent memory could be read and could contain data in which the intruder could be interested.
The problematic point is located in the following C-code in OpenSSL 1.0.1f:

```
int
tls1_process_heartbeat(SSL *s)
    {
    unsigned char *p = &s->s3->rrec.data[0], *pl;
    unsigned short hbtype;
    unsigned int payload;
    unsigned int padding = 16; /* Use minimum padding */

    /* Read type and payload length first */
    hbtype = *p++;
    n2s(p, payload);
    pl = p;

    if (s->msg_callback)
        s->msg_callback(0, s->version, TLS1_RT_HEARTBEAT,
            &s->s3->rrec.data[0], s->s3->rrec.length,
            s, s->msg_callback_arg);
```

It corresponds to the following sequence of instructions:

```
415d40:   48 8b 83 80 00 00 00    mov    rax,QWORD PTR [rbx+0x80]
415d47:   8b 90 24 01 00 00       mov    edx,DWORD PTR [rax+0x124]
415d4d:   31 c0                   xor    eax,eax
415d4f:   83 fa 12                cmp    edx,0x12
415d52:   76 2b                   jbe    415d7f <tls1_process_heartbeat+0x8f>
415d54:   44 0f b7 65 01          movzx  r12d,WORD PTR [rbp+0x1]
415d59:   66 41 c1 c4 08          rol    r12w,0x8
415d5e:   45 0f b7 ec             movzx  r13d,r12w
415d62:   45 8d 75 13             lea    r14d,[r13+0x13]
415d66:   44 39 f2                cmp    edx,r14d
415d69:   72 14                   jb     415d7f <tls1_process_heartbeat+0x8f>
415d6b:   0f b6 55 00             movzx  edx,BYTE PTR [rbp+0x0]
415d6f:   4c 8d 7d 03             lea    r15,[rbp+0x3]
415d73:   66 83 fa 01             cmp    dx,0x1
415d77:   74 47                   je     415dc0 <tls1_process_heartbeat+0xd0>
415d79:   66 83 fa 02             cmp    dx,0x2
415d7d:   74 11                   je     415d90 <tls1_process_heartbeat+0xa0>
415d7f:   48 83 c4 18             add    rsp,0x18
415d83:   5b                      pop    rbx
415d84:   5d                      pop    rbp
415d85:   41 5c                   pop    r12
415d87:   41 5d                   pop    r13
415d89:   41 5e                   pop    r14
415d8b:   41 5f                   pop    r15
415d8d:   c3                      ret
415d8e:   66 90                   xchg   ax,ax
415d90:   66 41 83 fc 12          cmp    r12w,0x12
415d95:   75 e8                   jne    415d7f <tls1_process_heartbeat+0x8f>
415d97:   0f b7 55 03             movzx  edx,WORD PTR [rbp+0x3]
```

This piece of code has been selected manually for demonstration and for proving of

vulnerability absence we should move from instruction 415d97 to instruction 415d40 by backward symbolic modelling. The initial state for backward moving is defined as the vulnerability case when length of buffer to be read is exceed the admissible one.
We define this state of vulnerability manually as the following formula:

Memory(rdi+128)<19 ||
Memory(rdi+128)>Memory(r12d)+19

In future, this formula would be derived from known semantics of memory allocation methods (malloc(), stack pointer, etc).
Since, this vulnerability was already resolved in OpenSSL 1.0.1g, we can compare how detection works. The older version does not contain input parameter checks so moving from vulnerability place to the initial state is possible and corresponding trace-counterexample has been generated. We will consider the second version of code where protection of the unauthorized access to the adjacent memory has been implemented. Providing symbolic modeling from vulnerability state we should not reach the initial state so the absence of vulnerability is proved.
Firstly, we need select the necessary slice of assembler code for symbolic modeling. It means that we will execute symbolically only those instructions that performs buffer access.
In the point of buffer access we have the following state of environment

(Memory(rdi+128)<19 || Memory(rdi+128)>
Memory(rol(8,Memory(rbp+1))+19)&
(rol(8,Memory(rbp+1))+19<=edx)&(edx>19)

This is disjunction of two states. If to consider every disjunction member we can see that they are not satisfiable, so the next basic protocols are non-applicable. We have deadlock by backward modeling so the absence of vulnerability is proved. The satisfiability of logical expressions was proved by Microsoft Z3 proving machine.

7 CONCLUSIONS

This simple example demonstrates the technology in a nutshell. This is the first step in big research that has started at the Glushkov Institute of Cybernetics together with Intel specialists.
The primary challenge is the formalization of the situation as a vulnerability formula under certain circumstances. Today it requires manual analysis of the code and vulnerability case, but it's possible to automate it. Another challenge is definition of the points where symbolic modeling should start for detection of the issues. There were two ways presented - automatic searching of possible suspected points in the code that present some input from external environment. The second way also could be implemented by symbolic modeling on a higher level of abstraction with computation of possible places where more detailed modeling should start, probably with using traces of actual execution.
Overall, the results are encouraging and demonstrate the perspective of the algebraic approach. The corresponding modules are developed in the scope of APS and will be updated to cover semantics of vulnerabilities exploited in the last virus attacks as an algebraic approach.

REFERENCES

[1] M. Kulenovic, D. Donko, A survey of static code analysis methods for security vulnerabilities detection, 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014.
[2] https://www.parasoft.com/product/static-analysis-cc/
[3] https://www.klocwork.com/
[4] https://clang.llvm.org/
[5] https://www.visualstudio.com/
[6] C.Pasareanu, Symbolic Execution of Java Byte code, ISSTA'08 Preceeding, 2008
[7] https://en.wikipedia.org/wiki/Fuzzing
[8] http://llvm.org/docs/LibFuzzer.html

[9] Algebraic Programming System APS www.apsystem.org.ua

[10] Common Vulnerabilities and Exposures (CVE®) https://cve.mitre.org/

[11] A. Letichevsky and D. Gilbert, "A Model for Interaction of Agents and Environments," in: *Resent trends in Algebraic Development technique, LNCS 1827* (D. Bert and C. Choppy, eds.), Springer-Verlag, pp. 311-328, 1999.

[12] A. Letichevsky, J. Kapitonova, O. Letychevskyi, V. Volkov, S. Baranov, V.Kotlyarov, T. Weigert, "Basic Protocols, Message Sequence Charts, and the Verification of Requirements Specifications," in: *Computer Networks*, pp. 662-675, 2005.

[13] Intel 64 and IA-32 Architectures Software Developer's Manual, Intel Corporation, 1997-2016

Context-Aware Service Discovery in the Internet of Things

Shiow-Fen Hwang, Guan-Yu Chen, and Chyi-Ren Dow
Department of Information Engineering and Computer Science, Feng Chia University
{sfhwang, m0312301}@mail.fcu.edu.tw, crdow@fcu.edu.tw

ABSTRACT

The internet of things(IoT) has rapid development in recent years, so that it is becoming more and more important in our lives. There will be more than billions devices connected to the internet in the near future. How to find the services (or resources) we needed in the internet is an important issue. There are many researches concerning the issue and a lot of novel schemes are proposed. However, most of them do not take into account the delay time and have high control overhead.

Therefore, in this paper, we propose an efficient context-aware service discovery mechanism that employs a spanning tree and has shorter delay time, less control packets and energy consumption in the internet of things. First, a service discovery spanning tree is established according to the weights of nodes defined by the number of services, residual energy and distance. Then, an efficient service discovery mechanism is presented by using Bloom filter and the established spanning tree. Moreover, a maintenance mechanism is provided for nodes joining or leaving. Simulation results indicate that the proposed method outperforms TRENDY and CAEsAR in terms of the number of packets and average search time.

Keywords: Internet of things, service discovery, context-aware, spanning tree, weight.

1 INTRODUCTION

With the rapid development of wireless communications technology, many devices are equipped with chips to connect to the Internet or communicate with each other, and thus form a Internet of things(IoT). There are many applications of IoT, such as manufacturing, medical and heath care, transportation, environmental monitoring, smart meter, smart home, smart city, etc.

The problem of service discovery is essential in IoT. However, due to the limited power capacity of nodes, how to find the services efficiently is important. Although there are a lot of studies concerning service discovery, most of them just proposed the architecture [1] [2] [3] [4]. For those studies that have proposed protocols [5] [6] [7] [8] [9] [10] [11] [12] [13] [14], some of them are centralized which need a powerful management center, and thus have a high cost; while the decentralized protocols rely on broadcasting to search for services, so lead to packet flooding and long search time. Hence, we design a hybrid service discovery protocol to avoid these disadvantages.

A. Kalmar et al. adapt RPL(IPv6 Routing Protocol for Low-Power and Lossy Networks) [15] to establish the route of data transmission for service discovery. They also use Bloom filter [16] [17] to compare the context information, and thus reduce the amount of compared data significantly. However, the route tree generated by RPL is not designed for the purpose of service discovery, so that it usually takes more search time and control packets. T.A. Butt et al. [9] propose a centralized context-aware service discovery protocol(TRENDY) which classifies nodes into four categories: Directory agent(DA), group members(GM), group leaders(GL) and user agents(UA). The DA stores and maintains the context information of all nodes. GMs send update messages periodically to DA. On top of GM's capability, GLs keep the record of assigned GMs by a list with their IP addresses and statuses. A UA sends a query to the DA for discovering service. RPL is applied as services have been discovered. Butt et al. [6] improve their previous works by using timer

and caching techniques. But there is still room for improvement.

Therefore, we propose a better service discovery method in the internet of things environments. Like TRENDY and CAEsAR, we adapt a tree architecture for discovering services. However, we define the weight of nodes by the number of services, residual energy and distance, so that the nodes with more services, high residual energy or smaller distance can be found early, and thus reduce the searching time.

2 CONTEXT AWARE SERVICE DISCOVERY

In this section, we propose an efficient service discovery and a maintenance mechanism in the internet of things networks.

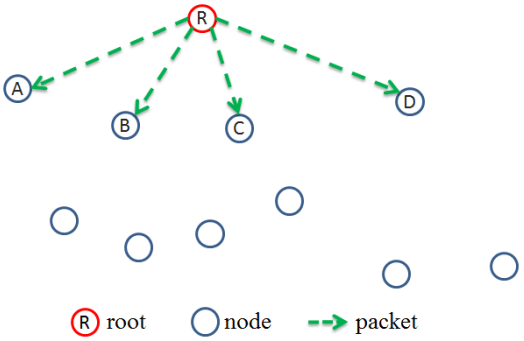
2.1 Service Discovery Spanning Tree

We choose the most powerful node in the network as the root to build a service discovery spanning tree. Then define the weight of each node *i*, *w(i)*, as follows.

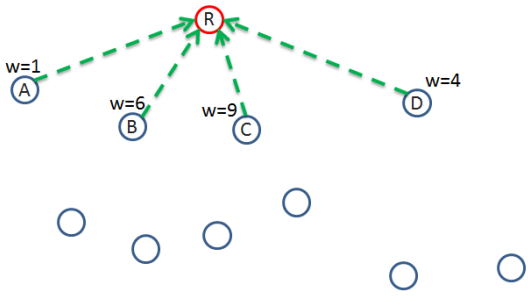
$$w(i) = \alpha \frac{s}{s_0} + \beta \frac{e_r}{e_0} + \gamma \frac{t_x}{2d} \qquad (1)$$

where $0 \le \alpha, \beta, \gamma \le 1$, $\alpha+\beta+\gamma=1$, s_0 is the average number of services of a node, s is the number of services provided by node i, e_0 is the initial energy of node i, e_r is the residual energy of node i, d is the distance from node i to the parent node, and t_x is the transmission range of node i.

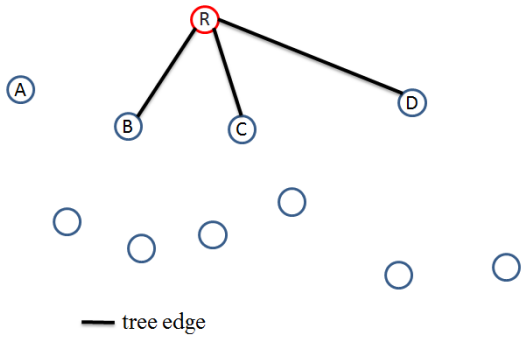
At the beginning, the root broadcast a message to its neighbor nodes. Every node received the message will return a packet with its weight to the root. The root informs the first k nodes of higher weight to be its children(if a node has no other neighbor or all its neighbors have already been in the current tree, it would have the highest priority to be one of the k selected nodes), and waits for their responses (the value k depends on the ability of the root and may be variable). If a node receives more than one packet that informs to be a child. It will choose the one with the highest weight as its parent, and record the others as its parent candidates. When a node *i* returns the confirmed message to its parent, the level of *i*, *l(i)*, is updated to *l*(its parent) plus one. (Initially, the level of the root is zero, and the levels of the other nodes are null.) At the same time, a part of the service discovering spanning tree is constructed. Afterwards, every leaf node in the current tree does the same process as its parent until the whole service discovering spanning tree is completed. After the service discovery spanning tree is constructed, every leaf node sends a packet including its level, linked list and Bloom filter which records the service information to its parent node. A parent node collects all such packets from its children and sends it to its parent. Repeat the same process until the root node. At this time, every node has the service information of its subtree, while the root has the service information of all nodes. Figure 1 is an example of the construction of a service discovery spanning tree.(Assume that the value k of the root is 3 and the others are 2)



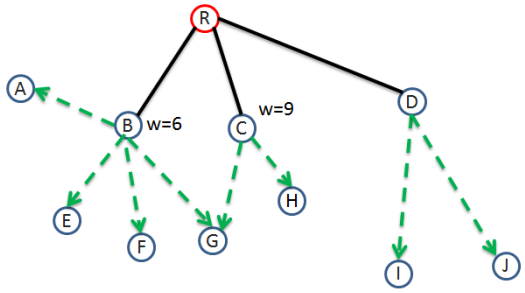
(a)



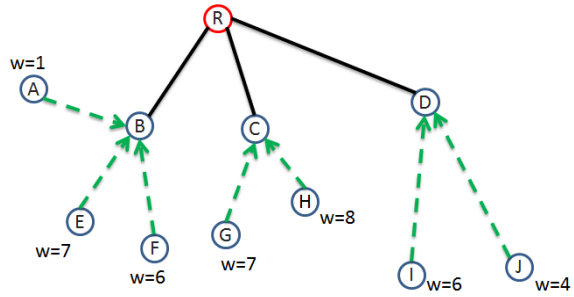
(b)



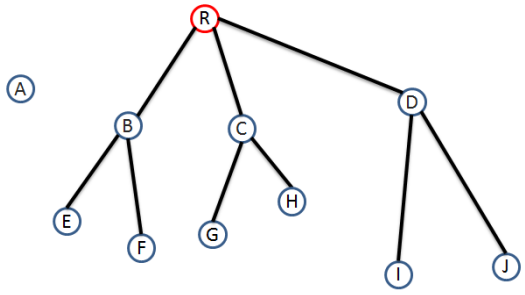
(c)



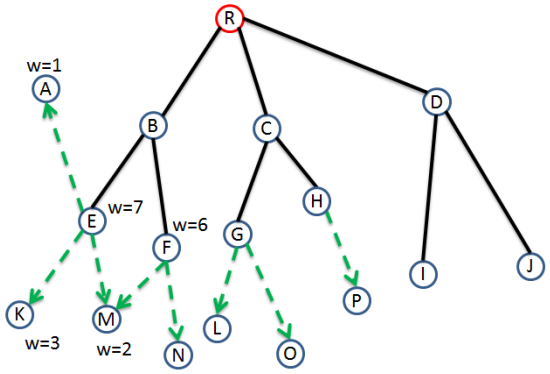
(d)



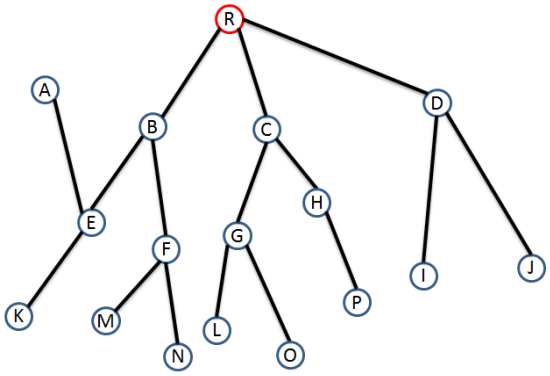
(e)



(f)



(g)



(h)

Figure 1. The construction of a service discovery spanning tree

2.2 Service Discovery Mechanism

With the service discovery spanning tree, if a node needs some services(resources), then it can query the root node. However, when the number of nodes is too large, there will be a numerous amount of stored data and a long searching time. In order to avoid this situation, we adapt linked lists and modify Bloom filter[16] to record service information. A Bloom filter is a bit array with predefined length and hash functions to represent a set of elements(services). Each service maps k bit positions of an array by k hash functions. And a linked list is established in the smallest of k bit positions to record the service name and the number of the service, as shown in Figure 2, the numbers of services S_1, S_2 and S_3 are 2, 1 and 1, respectively.

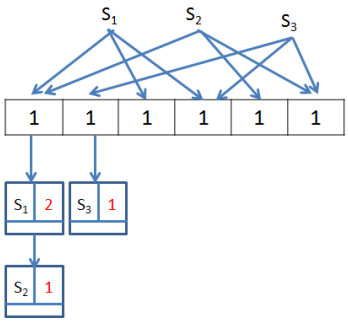


Figure 2. A Bloom filter with linked list

In our service discovering spanning tree, each node has Bloom filters of itself, left subtree and right subtree, and applies OR operation to get the union of these Bloom filters to send to its parent for saving space. When a node is requested for searching some services, if it can satisfy the required services, then the service discovery is done. Otherwise, it will check the Bloom filters of subtrees. If a subtree has the required services, it will go down to query its child. (If there are more than one child with required services, choose the one that the maximum level of its leaves is minimum. If tie break occurs, choose the one with maximum number of the required service and the highest weight in the order of priority.) Repeat the process until the request is satisfied or no required service exist. In Figure 3, the root R receives a query for service S_3. It checks the Bloom filter of right subtree, $W_{BF} \vee Z_{BF}$, (finds that service S_3 exists) and then sends the query to its child W. Finally, service S_3 is founded in Bloom filter Z_{BF} by W.

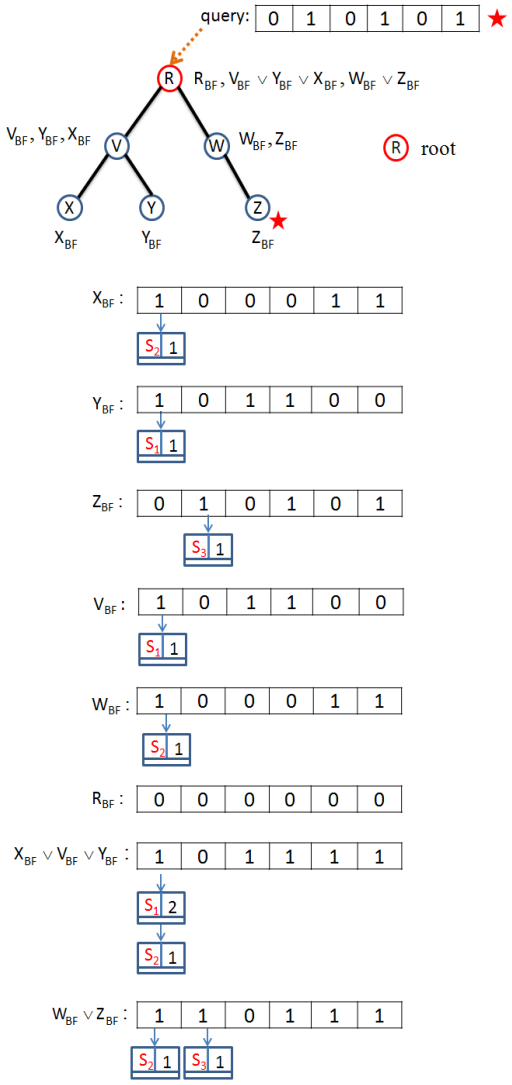


Figure 3. A Bloom filter with linked list for service discovery

Most of service discovery schemes assume that a service request comes from an external node and thus is sent to the root node directly. In fact, an internal node may also have a service request. So we also consider this case. When an internal node has a service request, it checks the Bloom filters of its subtree(if they exist) and searches for the required services. If the number of the required service in the subtrees is not enough, it will query its parent for an insufficient amount of required services. If it is still not enough, the node will query its candidate parents first and then its grandparent until the root if necessary. Figure 4 and Figure 5 are examples of our service discovery mechanism

in the cases of external request and internal request, respectively.

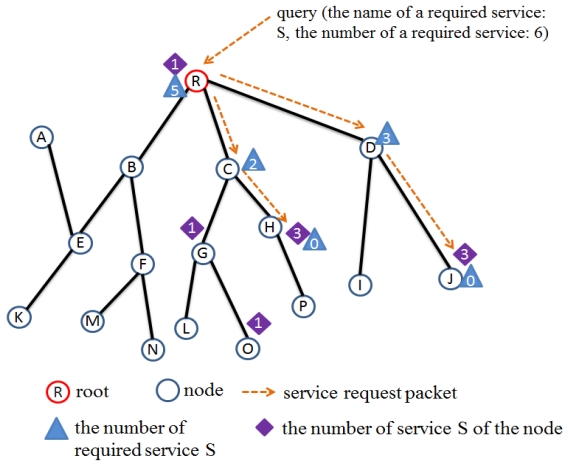


Figure 4. An external request for service discovery

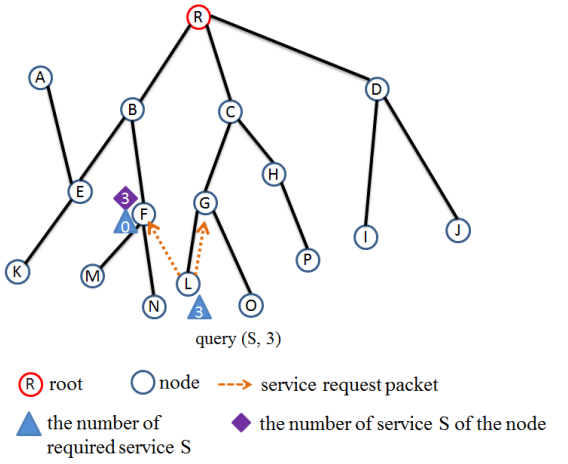
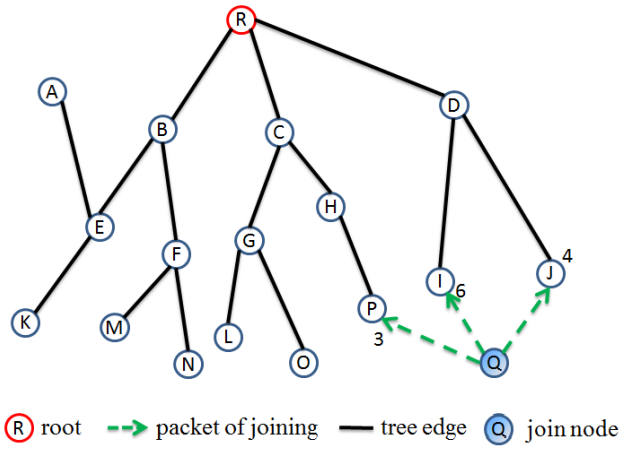


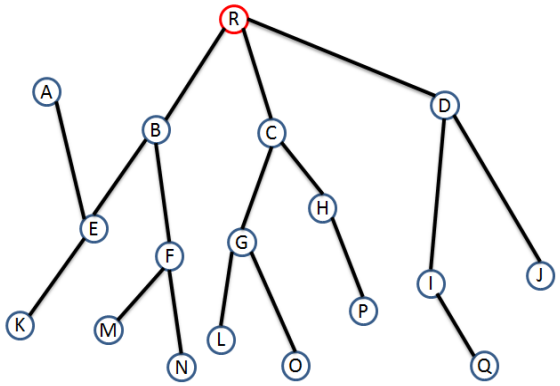
Figure 5. An internal request for service discovery

2.3 Maintenance

If a new node would like to join the service discovery spanning tree, it will broadcast a message to neighbors. Those which receive the message and are able to have more children return messages with their weights and levels. The new node choose the one with the smallest level(or the maximum weight if tie break) as its parent, and sends a confirmed message including its weight, level and services to its parent. After receiving the message, all information stored in its ancestors have to be updated. Figure 6 is an example of the maintenance for node joining.



(a)

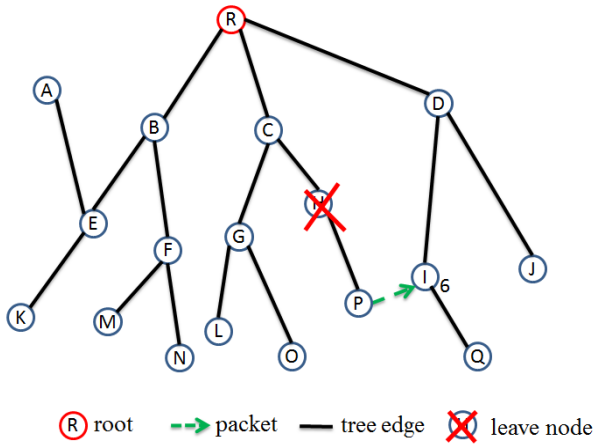


(b)

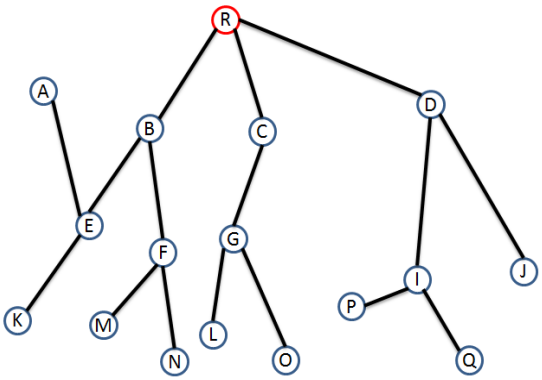
Figure 6. The maintenance for node joining

On the contrary, when a node would like to leave the service discovery spanning tree, if it is a leaf node, it just needs to send a message to its parent for updating the information of all its ancestors. Otherwise, it also needs to inform its children, and then they are going to look for new parents as the process of node joining. Figure 7 is an example of the maintenance for node leaving.

ISBN: 978-1-941968-42-0 ©2017 SDIWC



(a)



(b)

Figure 7. The maintenance for node leaving

3 SIMULATION RESULTS AND ANALYSIS

We evaluate the performance of the proposed mechanism and compare to TRENDY[9] and CAEsAR[10] according to the different transmission ranges and number of services of nodes in terms of the average search time and number of packets.

The simulator Cooja and OS Contiki are used for simulation. The simulation environment size is 50m x 100m to 100m x 200m. The number of services of a node is 2 to 6, and the average number is 3. Table 1 is the simulation parameters.

Table 1. simulation parameters

Parameters	Value
Transmission Range	30m、40m、50m
Number of Nodes	10、15、20、25、30
α、β、γ	0.5、0.2、0.3

Figure 8 to Figure 10 show the number of packets for all schemes when the transmission range are 30m, 40m and 50m, respectively. The results indicate that the proposed scheme has less number of packets than the others. This is because the proposed scheme supports multiple service search by modified Bloom filter and linked list, and usually searches the services near the requester by the use of level.

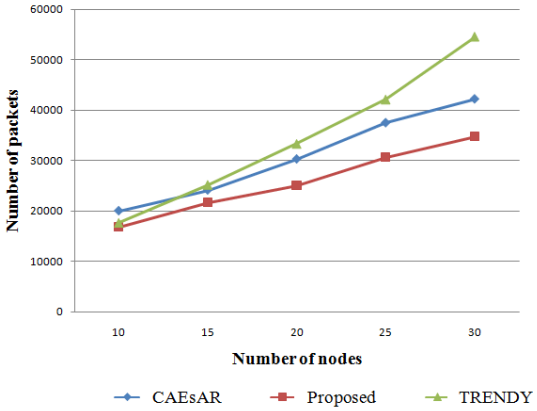


Figure 8. Number of packet (transmission range=30m)

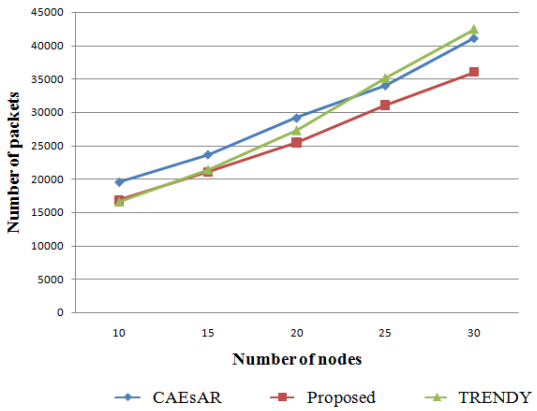


Figure 9. Number of packet (transmission range=40m)

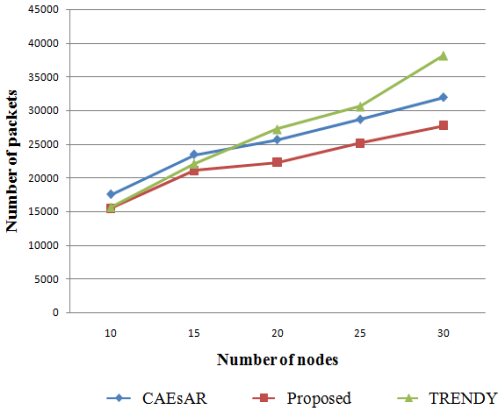


Figure 10. Number of packet (transmission range=50m)

Figure 11 to Figure 13 are the impacts of the number of nodes versus the average search time under the transmission ranges of 30m, 40m and 50m, respectively. The average search time increases obviously as the number of nodes increase in CAEsAR, while the proposed scheme and TRENDY increase slowly, especially under transmission ranges of 40m and 50m. This is because the distance and the number of services are taken into account for constructing the service discovery spanning tree in the proposed scheme, and TRENDY is centralized.

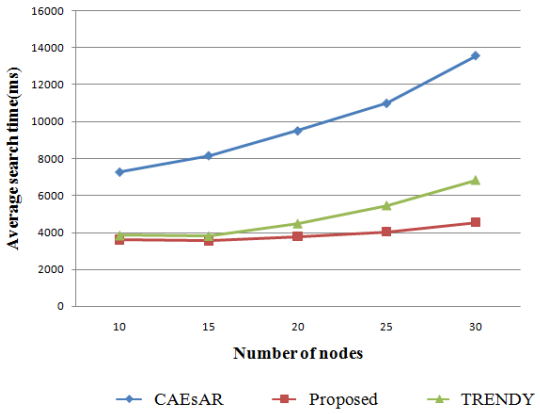


Figure 11. Average search time (transmission range=30m)

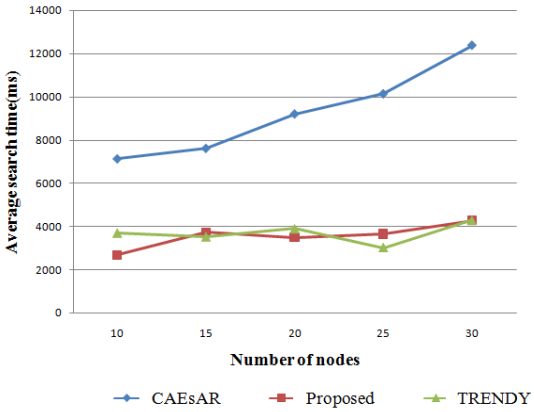


Figure 12. Average search time (transmission range=40m)

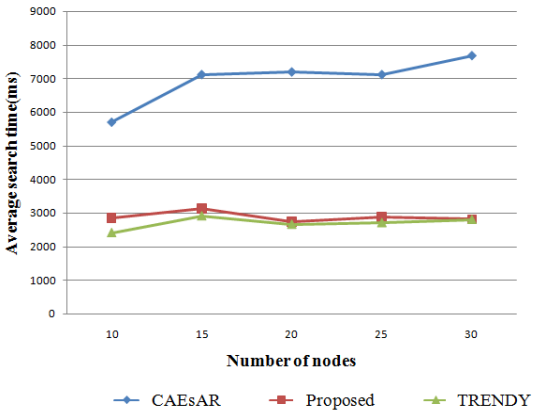


Figure 13. Average search time (transmission range=50m)

4 CONCLUSIONS

By establishing the service discovery spanning tree and modifying Bloom filter, we propose a context-aware service discovery mechanism in the internet of things. Simulation results show that our proposed mechanism has less number of packets than TRENDY and CAEsAR, and also has shorter search time than CAEsAR.

ACKNOWLEDGMENTS

This work was partially supported by the Ministry of Science and Technology, Taiwan, R.O.C. under Grants MOST-105-2221-E-035-059-.

REFERENCES

[1] S.A. Chaudhry, W.D. Jung, C.S. Hussain, A.H. Akbar, and K.H. Kim, “A Proxy-Enabled Service Discovery Architecture to Find Proximity-Based Services in 6LoWPAN,” EUC'06 Proceedings of

the 2006 international conference on Embedded and Ubiquitous Computing, vol. 4096, pp. 956-965, August 2006.

[2] J. Leguay, K. Jean-Marie, M. Lopez-Ramos, and V. Conan, "An Efficient Service Oriented Architecture for Heterogeneous and Dynamic Wireless Sensor Networks," 2008 33rd IEEE Conference on Local Computer Networks (LCN), pp. 740-747, October 2008.

[3] A. Jara, P. Lopez, D. Fernandez, J.F. Castillo, M.A. Zamora, and A.F. Skarmeta, "Mobile Digcovery: A Global Service Discovery for the Internet of Things," 2013 27th International Conference on Advanced Information Networking and Applications Workshops, pp. 1325-1330, March 2013.

[4] S. Cirani, L. Davoli, G. Ferrari, R. Léone, P. Medagliani, M. Picone, and L. Veltri, "A Scalable and Self-Configuring Architecture for Service Discovery in the Internet of Things," IEEE Internet Of Things Journal, vol. 1, no. 5, pp. 508-521, September 2014.

[5] D. Parlanti and F. Paganelli, "A DHT-Based Discovery Service for the Internet of Things," Journal of Computer Networks and Communications, vol. 2012, pp. 1-11, September 2012.

[6] G. Oikonomou, I. Phillips, L. Guan, and T. A. Butt, "Adaptive and Context-Aware Service Discovery for the Internet of Things," 13th International Conference, NEW2AN 2013 and 6th Conference, ruSMART 2013, vol. 8121, pp. 36-47, August 2013.

[7] A. Kovacevic, J. Ansari, and P. Mahonen, "NanoSD: A Flexible Service Discovery Protocol for Dynamic and Heterogeneous Wireless Sensor Networks," 2010 Sixth International Conference on Mobile Ad-hoc and Sensor Networks, pp. 14-19, December 2010.

[8] D. Schoder, K. Fischbach, and N. Schoenemann, "P2P architecture for ubiquitous supply chain systems," 17th European Conference on Information Systems, pp. 2255-2266, June 2009.

[9] G. Oikonomou, I. Phillips, L. Guan, and T. A. Butt, "TRENDY: An Adaptive and Context-Aware Service Discovery Protocol for 6LoWPANs," WoT '12 Proceedings of the Third International Workshop on the Web of Things, pp. 4-9, June 2012.

[10] A. Kalmár, M. Maliosz, and R. Vida, "CAEsAR : A Context-Aware Addressing and Routing Scheme for RPL Networks," 2015 IEEE International Conference on Communications (ICC), pp. 635-641, June 2015.

[11] G. Min, L. Liu, R. Chen, and Z. Li, "Dynamic Resource Discovery based on Preference and Movement Pattern Similarity for Large-Scale Social Internet-of-Things," IEEE Internet Of Things Journal, vol. 3, no.4, pp. 581-589, June 2015.

[12] B. Djamaa and R. Witty, "An Efficient Service Discovery Protocol for 6LoWPANs," 2013 Science and Information Conference, pp. 645-652, October 2013.

[13] A. ElMougy and R. Helal, "An Energy-Efficient Service Discovery Protocol for the IoT based on a Multi-Tier WSN Architecture," 2015 IEEE 40th Local Computer Networks Conference Workshops, pp. 862-869, October 2015.

[14] D. Guinard and S. Mayer, "An Extensible Discovery Service for Smart Things," WoT '11 Proceedings of the Second International Workshop on Web of Things, pp. 7, June 2011.

[15] T. Winter, P. Thubert, A. Brandt, J. Hui, R. Kelsey, P. Levis, K. Pister, R. Struik, J. Vasseur, and R. Alexander, "RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks," RFC 6550 (Proposed Standard), March 2012. [Online]. Available: http://www.ietf.org/rfc/rfc6550.txt.

[16] B.H. Bloom, "Space/Time Trade-Offs in Hash Coding with Allowable Errors," Communications of the ACM, vol. 13, no. 7, pp. 422-426, July 1970.

[17] C.K. Chang, L.J. Zhang, and S. Cheng, "An Efficient Service Discovery Algorithm for Counting Bloom filter-Based Service Registry," 2009 IEEE International Conference on Web Services, pp.157-164, July 2009.

Improvement of Load Balancing Method in a Distributed Web System Using DNS

Kota MORIGAKI and Keizo SAISHO
Kagawa University
2217-20 Hayashi-cho, Takamatsu 761-0396, Japan
s16g473@stu.kagawa-u.ac.jp, sai@eng.kagawa-u.ac.jp

ABSTRACT

Progress of virtualization technology in recent years made it easy to build virtual servers on cloud. They can be used as cache server for load balancing. However, expected responsiveness cannot be gained with insufficient cache servers against load. In contrast, costs will increase by surplus cache servers against load. Therefore, we have been developing a distributed web system that adjust the number of cache servers according to load of them to reduce running cost. In this study, a load balancing method using DNS round-robin is now developed. However, load imbalance occurs among the servers with this method and responsiveness decreases because it is difficult to distribute the load uniformly using DNS round-robin. Therefore, we implement a function to suspend the allocation of requests to the overloaded server. This paper describes improvement of load balancing method and evaluation of it. From results of experiments, we confirm that improved function is possible to prevent lowering responsiveness with lower TTL value.

KEYWORDS

DNS Round-Robin, Distributed Web System, Load Balancing, Suspend Allocation, Auto-scaling, Cache Server

1 INTRODUCTION

In recent years, the Internet users increase and much service is performed using Web. Therefore, load of Web servers is growing more and more. If the load is over the limit of server's capacity, it returns the response with large delay, and it goes down in the worst case. Load balancing techniques are often used to avoid overload of servers. There is a Web system that distributes requests to multiple servers such as cache servers or mirror servers for load balancing. Progress of virtualization technology made it easy to build virtual servers on cloud. They can be used as cache server. Responsiveness, however, does not improve with insufficient cache servers against load. In contrast, costs will increase with surplus cache servers against load. Therefore, we developed a distributed web system using Load Balancer that dynamically adjust the number of cache servers according to load of them to reduce running cost[1]. In this study, we develop a method to distribute load to cache servers using DNS round-robin. However, load imbalance occurs among the servers with this method and responsiveness decreases because it is difficult to distribute the load uniformly using DNS round-robin. Therefore, we implement a function to suspend the allocation of requests to the overloaded server.

2 RELATED WORKS

A cloud auto-scaling mechanism aiming at providing necessary resources at low cost has been studied[2][3]. In [2], auto-scaling mechanism based on workload information and performance desire is implemented in Windows Azure platform. The result of the experiment shows that cost can be reduced by choosing an instance type of appropriate performance for the workload. This research covers a variety of applications, but our system targets only web application and aims at cross-use multiple cloud services.
In [3], an auto-scaling algorithm based on the number of active sessions of the web server is described. A load balancer is used for load balancing. Although we also use the number of active sessions as the load value, we use DNS for load balancing.

In [4], dynamic load balancing method using dynamic DNS update and round-robin mechanism is proposed. In this method, a server is dynamically added to or removed from the DNS list. The scheduling algorithm considers usage rates of server's CPU, memory, and network. The result of the experiment shows that both the response time and the average file transfer rate of the proposed system are faster than those of a pure round-robin DNS. It is similar to our load balancing method, but ours uses the number of active sessions as the load value.

3 DISTRIBUTED WEB SYSTEM USING DNS

Figure 1 shows our distributed Web system using DNS which consists of management server, authoritative name server, origin server and cache servers on cloud. The origin server services original contents and cache servers service the cache of them. The managing server manages the number of cache servers and DNS zone of the authoritative name server. For load balancing, this system uses DNS round-robin method which sends the list of IP addresses in a different order to a new client each time. Most clients use the first IP address they receive to connect server. Therefore, requests from clients are sent to each server. By managing the DNS zone of the authoritative name server, it is possible to control the start and stop of allocating request to each server.

The management server has the following functions.

- Load monitoring function
  The load monitoring functions monitors load of the origin server and cache servers. This function periodically measures the current and the maximum number of Web server processes and calculates ratio of the current number against the maximum number (Operating Ratio), and calculates average of Operating Ratio of working servers (Average Operating Ratio, AVGOR). This system uses AVGOR as load value.
- Cache server management function
  The cache server management function boots up and shuts down cache servers. This function decides the number of required cache servers based on AVGOR obtained by the load monitoring function. When AVGOR is greater than threshold of scale-out (Th_{high}), it boots up a new cache server. When AVGOR is less than threshold of scale-in (Th_{low}), it shuts down a latest booted cache server.
- DNS management function
  The DNS management function manages the DNS zone of authoritative name server. According to the booting up and shutting down cache server, IP address of it is added to or removed from the DNS zone dynamically. When the load monitoring function cannot monitor load of a server, it also removes the server. Therefore, it is possible to cope with server failure such as system down.

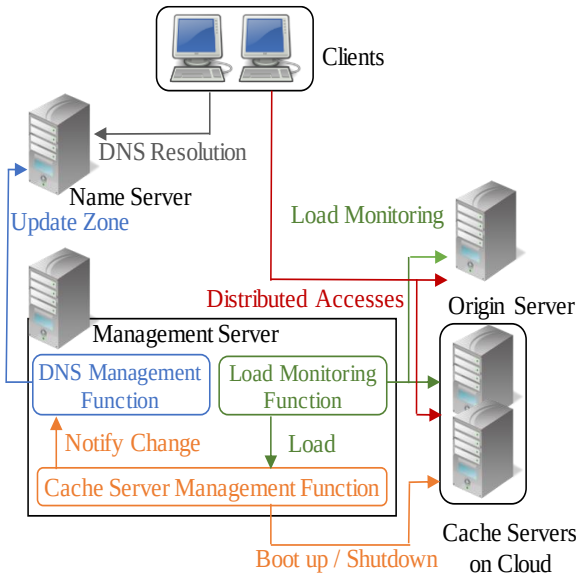


Figure 1. Distributed Web System using DNS

4 LOAD BALANCING USING DNS ROUND-ROBIN

We experimented with the distributed Web system described in the previous section. The result showed, load imbalance among the servers using the DNS round-robin method.

4.1 Experiment Environment

Figure 2 shows the experiment environment. All servers and clients are built as virtual machine on hypervisors which specifications are shown in Table 1. The management server and DNS servers are built on hypervisor1, the origin server and nine cache servers are built on hypervisor2 and hypervisor3, and twelve clients are built on hypervisor4. Mirror server is used instead of cache server because cache mechanism is now developing. Apache2.4[5] is used as a Web server software. DokuWiki[6] runs on all servers. Each client accesses the web server using Siege[7]. Siege is the stress test tool. The number of simultaneous accesses is set to 100. Therefore, the maximum number of simultaneous accesses is 1,200 (100×12). TTL value for DNS is set to 60 seconds. Th_{high} and Th_{low} are set to 0.6 and 0.1, respectively.

4.2 Experiment Procedure

The scenario of the experiment is shown in below. To examine the load and the response time of each server, the number of simultaneous accesses to web servers is stepwise changed.

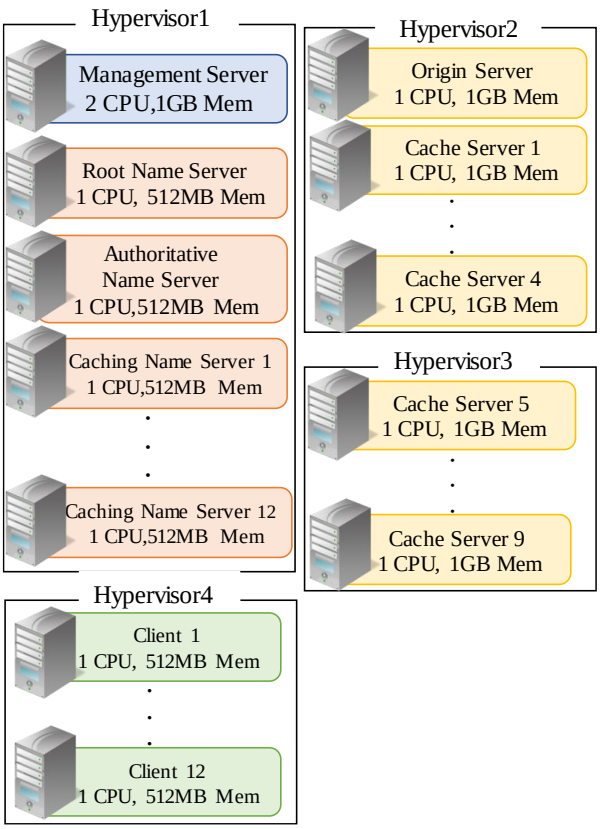


Figure 2. Experiment environment

Table 1. Spec of each hypervisor

	CPU	Memory
Hypervisor1	Intel Xeon E5-2620	32GB
Hypervisor2	Intel Xeon E5-2620	32GB
Hypervisor3	Intel Xeon E5-2620	32GB
Hypervisor4	Intel Core i7-4790K	32GB

I. Start with no accesses.
II. Add 1 client every 30 seconds.
III. After all clients are added, keep all clients accessing for 500 seconds.
IV. Remove 1 client every 30 seconds.
V. End when no accesses.

4.3 Experiment Result

Figure 3 shows Operating Raito of each server. Figure 4 shows the response time and the Operation Raito of the origin server.
In Figure 3, several servers are overloaded for a long time, and some servers remain low load. This phenomenon happens clearly around 500 seconds.
In Figure 4, purple line and blue line show average response time and maximum response time for one second, respectively, green line shows Operation Ratio. Maximum response time varies very much. Requests from clients are distributed to each server by using the round-robin method. However, this method cannot consider the load of each server and it causes load imbalance among the servers and response time lengthens.

5 SUSPENDING FUNCTION

We think that the problem described in the previous section can be coped with by suspending the allocation of requests to the overloaded server. We implement a function that excludes overloaded servers from DNS answer. We call this function suspending function. The function uses the PipeBackend of PowerDNS[8] that is DNS software. It can call external program that resolves DNS queries dynamically through PipeBackend module. We implement the program that resolves DNS queries based on the

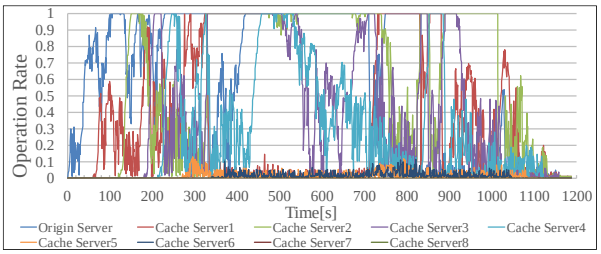


Figure 3. Operation Ratio of each server

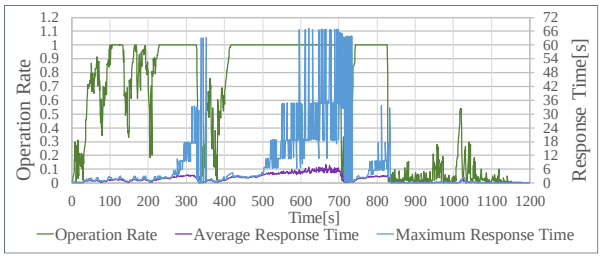


Figure 4. Operation Ratio and response time of origin server

configuration file shown in Figure 5. The file contains a hostname, IP address, status value of each server and TTL value. The status value is updated every second based on Operation Ratio of each server obtained by the management server and sent to the authoritative name server. The status value is set to 0 while the corresponding server stay in overloaded. otherwise, the status value is set to 1. The IP address is included in DNS answer if the corresponding status value is 1. In contrast, the IP address is excluded from DNS answer if the corresponding status value is 0. For example, IP address 192.168.11.21 and 192.168.11.22 is included in DNS answer with configuration shown in Figure 5.

6 EVALUATION

In this section, we evaluate the function described in the previous section. The experiment environment and the experiment procedure are the same as in Section 4. When the Operation Ratio of the server is the threshold value and over, the function decides that the server is overloaded and set the status value in configuration file to 0. Otherwise, the value is set to 1. In this experiment, the threshold value is set to 0.6 (case A), 0.8 (case B) or 1 (case C). Experiment without the

```
example.com:
  A:
    IP:
        192.168.11.21: 1
        192.168.11.22: 1
        192.168.11.23: 0
        192.168.11.24: 0
        192.168.11.25: 0
        192.168.11.26: 0
    TTL: 60
```

Figure 5. A Sample configuration file

suspending function is represented as case D. We performed experiments ten times in all cases. In order to investigate influence of the function to suspend allocation of requests to overloaded servers, the experiment results while number of simultaneous accesses is maximum are examined.

The average of the results is shown Table 2. The cost is sum of uptime of all servers. The average response time in case D is the best in all cases. The number of requests per cost in case D is also the best.

Table 2. Experiment result of each case (TTL 60)

Case	Threshold value	Cost	Total response time	Total number of requests	Average response time	number of requests per cost
A	0.6	3454	445152	312404	1.42	90.45
B	0.8	3316	447405	309476	1.45	93.32
C	1	3279	454476	292067	1.56	89.08
D	Without function	3258	431490	304978	1.41	93.60

Table 3 shows the percentage of access with response time longer than or equal to 3,9,15,30 and 60 seconds. The blue and red letters indicate the lowest and highest percentage, respectively. Percentages of access with response time longer than equal to 3 and 60 seconds severally in case D are the highest. In contrast, other percentages in case D are the lowest.

Table 3. Percentage of long responded access (TTL 60)

Case	Threshold value	3	9	15	30	60
A	0.6	15.103	2.453	1.269	0.468	0.031
B	0.8	14.760	2.910	1.569	0.609	0.027
C	1	17.070	3.068	1.692	0.656	0.030
D	Without function	18.671	1.403	0.650	0.262	0.074

The results show the function is ineffective. In DNS, the TTL value specifies the expiration date of the DNS cache. It takes long time to reflect the updated DNS zone with large TTL.

ISBN: 978-1-941968-42-0 ©2017 SDIWC

Therefore, the suspending function takes no effect on response time. So, the same experiments except TTL value set to 30 seconds is performed.
The average of results is shown Table 4. The average response time in case C is the best in all cases. The number of requests per cost in case C is also the best.

Table 4. Experiment result of each case (TTL 30)

Case	Threshold value	Cost	Total response time	Total number of requests	Average response time	number of requests per cost
A	0.6	3745	417580	368184	1.13	98.33
B	0.8	3712	419933	364384	1.15	98.16
C	1	3627	416003	371359	1.12	102.40
D	Without function	3407	433705	320231	1.35	93.99

Table 5 shows the Percentage of long responded access. All percentages of access in case A are the lowest in all cases. The percentage of access with response time longer than equal to 3 seconds in case D is the worst in all cases and more than twice as high as that in case A.
Figure 6 and Figure 7 show the response time and the Operation Raito of the origin server in case A and case D, respectively. Line colors are same as in Figure 4. In Figure 6, the maximum response time is about half of the maximum response time in Figure 7. These results show the effectiveness of the suspending function with small TTL.

Table 5. Percentage of long responded access (TTL 30)

Case	Threshold value	3	9	15	30	60
A	0.6	8.567	0.817	0.400	0.103	0.0000
B	0.8	9.160	0.936	0.461	0.118	0.0000
C	1	10.268	1.009	0.442	0.122	0.0004
D	Without function	19.164	1.011	0.425	0.147	0.0360

7 CONCLUSION

We implemented the suspending function to exclude overloaded servers from DNS answer and evaluated it. By the experiment, it is confirmed that the function is possible to improve responsiveness with lower TTL value. However, the function reduces responsiveness with higher TTL value.
The followings are future works.

- Practical Scenarios of Experiment
- Examination of TTL value
- Experiment using cloud environment

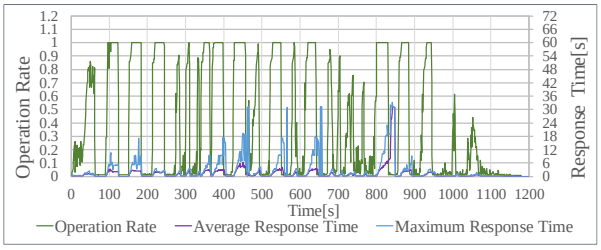


Figure 6. Operation Ratio and response time of origin server in case A (TTL 30)

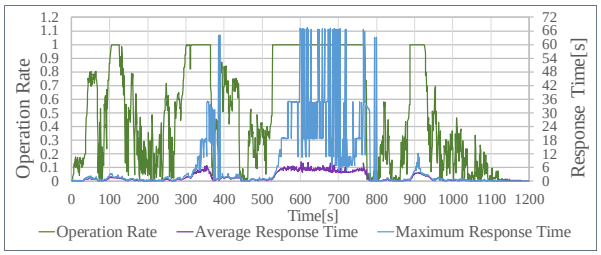


Figure 7. Operation Ratio and response time of origin server in case D (TTL 30)

REFERENCES

[1] A. Horiuchi, and K. Saisho. “Prototyping and Evaluation of Virtual Cache Server Management Function for Distributed Web System,” The 2015 International Conference on Computational Science and Computational Intelligence (CSCI'15), pp.324-329, 2015.

[2] M. Mao, J. Li, and M. Humphrey. “Cloud auto-scaling with deadline and budget constraints,” 11th ACM/IEEE International Conference on Grid Computing (Grid 2010), 2010.

[3] T.C. Chieu, A. Mohindra, A.A. Karve, and A. Segal. “Dynamic Scaling of Web Applications in a Virtualized Cloud Computing Environment,” 2009 IEEE International Conference on e-Business Engineering, pp.281-286, 2009.

[4] JB. Moon, and MH. Kim, “Dynamic Load Balancing Method Based on DNS for Distributed Web Systems,” International Conference on Electronic Commerce and Web Technologies, pp. 238-247, 2005.

[5] Apache, https://httpd.apache.org/

[6] DokuWiki, https://www.dokuwiki.org/dokuwiki#

[7] Siege, https://www.joedog.org/siege-home/

[8] PowerDNS, https://www.powerdns.com/

Survey on Open Source Frameworks for Big Data Analytics

Chao-Hsu Chen and Chien-Lung Hsu
Department of Information Management, Chang Gung University,
No. 259, Wenhua 1st Rd., Guishan Dist., Taoyuan City 33302, Taiwan
E-mail: M0345208@stmail.cgu.edu.tw, clhsu@mail.cgu.edu.tw
Kuo-Yu Tsai
Department of Applied Mathematics, Chinese Culture University,
No. 55, Hwa-Kang Road, Yang-Ming-Shan, Taipei 11114, Taiwan
E-mail: cgy13@ulive.pccu.edu.tw

ABSTRACT

The main contribution of this study is to provide a feasible open source big data framework and comparison of characteristics of modules upon different layers so that whenever big data processing technologies is essential for industries and business applications to deploy with. In accord with comparison to successful models, open source big data framework and modules with different characteristic can be quickly referred to and adopted. Based upon this study, we can develop specific open source big data framework and modules for different industries in future.

KEYWORDS

Open Source, Big Data, Framework

1 INTRODUCTION

To the huge amount of growing data, Google has developed Google File System [1] and MapReduce [2] and also developed a variety of open-source software system. This leads to the development of large data handling software like Apache Hadoop [3] and Hadoop File System [4]. Big data services and solutions are experiencing a sustained growth, which also reflect the growing number of major vendors such as IBM [5], Oracle [6], HPE [7], Microsoft [8], SAS [9] and Cloudera [10].

Thus, comparative analysis and benchmarking big data platform has become increasingly important. There are many well-known large companies hunting for big data talented people and resources to construct large data applications [11, 12]. They do can take advantage of their huge corporate data to make business strategies faster, more efficient, reducing the resources and time-consuming, and thus enhance their competitiveness. In the construction of a complete big data applications, usually in data collection, storage, management, processing, rendering visualization, privacy control, business model and other sectors, it needs experts and expertise to resolve. Therefore, data, domain expert and communication specialists also appears to play an important role in big data era.

Big data has evolved through volume, variety, velocity, 3V [13] to volume, variety, velocity, value, veracity, 5V [14] in big data analysis capabilities. Many companies have faced one-day amount of data at a rate of tens, hundreds, from TB (terabyte) increases to PB (petabyte) level, so that the traditional database becomes difficult to handle the amount of data whereas volume is an important factor in these days.

Velocity is important since data nowadays is increasing faster and faster, such as mobile computing. Along with the popularity of social networks, data increases faster than traditional enterprise applications. With the flow of faster information, data processing and analysis have to speed up to keep everything up. Variety refers to the diversity of information, the internet is now not only the information, apart from information we post pictures, videos, and backup data. Value

means the ability to manage these increasing big data ecosystem. Veracity is relatively important since the correctness and accuracy of the information are very important these days. In this new era, big data is the heart of all the information. Many vendors rely on the analysis of their new generation systems to achieve and deal with this huge data.

In this paper, we will have a clearer understanding of the advantages and disadvantages of each open source platform usage of big data platform in the industry. We can save more time in referencing analytical framework. Finally, by the trend of big data platforms by international development to make useful recommendations of large data analysis applications.

2 BIG DATA INFRASTRUCTURE

2.1 Hadoop Stack Analytics

Hadoop platform as shown in the following two most important members as shown in Figure Figure 1 Hadoop Stack with different components. Hadoop Distributed File System (HDFS) [15] is to store data across a cluster of machines providing high availability and fault tolerance of distributed file system. Hadoop YARN [16] handles resource management and scheduling job across the cluster.

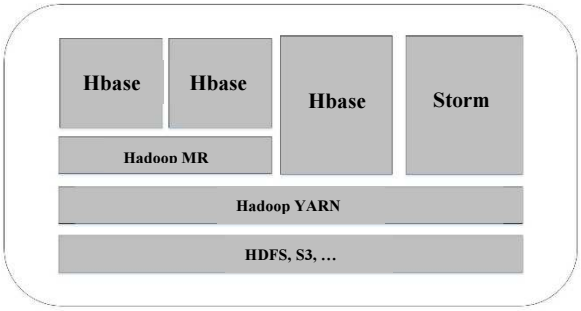


Figure 1 Hadoop Stack with different components

- ***MapReduce***

MapReduce [17] programming model used in Hadoop was proposed by Google's Dean and Ghemawat. The basic scheme in Hadoop MapReduce processing can be divided into two parts, known as mappers and reducers. At a high level, mappers read out data from HDFS to be processed, and produce results to reducers. Reducers are used in the polymerization intermediate results to produce the final output which is written to HDFS again. A typical Hadoop job involves running on different nodes in the multiple cluster of mappers and reducers. A good use of MapReduce can be found out in parallel data processing [18].

- ***MapReduce wrappers***

A specific packaging (Wrapper) for MapReduce is currently being developed. These packages provide better control of the source code to MapReduce to assist with.

- **Apache Pig:** Yahoo [19] developed a SQL-like environment being used in many organizations such as Yahoo , Twitter , Facebook, AOL, LinkedIn, etc.
- **Hive:** Facebook [20] developed another MapReduce packaging. Both packages provide a better environment in which the program can develop easier where the app developers do not have to deal with complex MapReduce program code.

- ***Limitations of MapReduce***

One of the major disadvantages of the MapReduce algorithm is its efficiency reduced when it is running iterative formula. Mapper [12] repeatedly reading the same data from the hard drive. The results of must be able to write to your hard drive each time before they are passed to the next. This bottleneck reduces MapReduce performance.

2.2 Berkeley data analytics stack (BDAS)

Spark developers also proposed what is known as a complete Data Processing Stack called Berkeley Data Analytics Stack (BDAS) [12, 21,] as in Figure **錯誤! 所指定的樣式的文字不存在文件中。**2. At the bottom of this stack , there is Tachyon [22] which is of HDFS component. The main advantage of Tachyon than Hadoop HDFS is that it can have more

aggressive memory usage. Another feature of Tachyon is its compatibility with Hadoop MapReduce. MapReduce can be run on Tachyon without any modification.

In BDAS, Tachyon upper layer is Apache Mesos. As a resource isolation and sharing to distributed applications in cluster manager, it has support to Hadoop, Spark, Aurora [23], and other applications. Mesos on scalability can be increased up to tens of thousands of nodes. In BDAS architecture, the third component running on the Mesos, Spark is part of playing the role of Hadoop MapReduce.

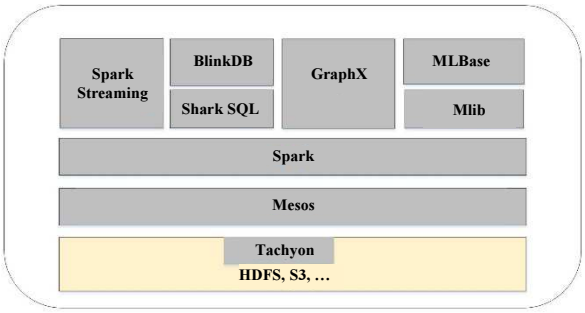


Figure 錯誤! 所指定的樣式的文字不存在文件中。 Berkeley Data Analysis Stack (BDAS) components

3 OPEN SOURCE BIG DATA PLATFORMS PRELIMINARY APPROACH

In this paper, a variety of open-source platform will massively discussed regarding scalability, data processing IO performance, data size, iterative support, real time processing and other advantages and disadvantages [11, 12]. Upon data collections and consulting domain experts, Hadoop Big Data Stack [15] and Berkeley data analytics stack (BDAS) [12, 21] different components and different data layer as in below table. There is fewer discussion regarding visualization and analysis at below table in BDAS.

Table 1 Big Data Infrastructure

	Open Source Hadoop Stack	BDAS
B.I.	Pentaho, SAP Cognos	Pentaho, SAP Cognos
Visualization	EDW, SAS, SAP, Qlikview, Tableau, D3.js, Kibana, Datawatch	
Analysis	EDW Connector, EDW, Hive, Impala, Stinger, Drill, Mahout, Mlib	Mlib
Data Mgmt	Oozie, Chukwa, Flume, Flumetd, Zookeeper, Spark SQL, Scribe, Logstash, Elasticsearch	Spark Streaming, Spark SQL, Blink DB, Graph X, ML Base
Data Access	Sqoop, Hive, Pig, Hbase, Apache Storm, Avro	Hadoop Map Reduced, Spark
Data Processing	Hadoop Map Reduced, Hadoop Yarn	Mesos
Data Storage	HDFS, S3, Hbase	HDFS, S3, Hbase, Tachyon

3.1 Characteristics and Discussion

With the explosive growth of information, more and more enterprises are deploying a private cloud system or hire a public cloud system to handle large data. Doug Cutting, creator of both Lucene and Hadoop, and Mike Cafarella originated Nutch and Hadoop history. Following is history of Hadoop creation [10].

Currently ways of analyzing big data platform can be classified into the following four [24, 25]：

- Transaction type RDBM Systems
  - **Enterprise Hubs:** IBM [26], Oracle [6], and Sybase [27] has traditional enterprise application analysis and processing transactions.
  - **Departmental Marts:** Microsoft SQL [8] and MySQL are suitable for small or medium business (SMB).

- Analytic Platforms
  - **MPP Database:** Massively parallel processing (MPP) data base providing vendors are Teradata [28], Microsoft[8] and EMC Greenplum [4].

- **Analytical Appliance:** IBM Netezza [26], EMC Greenplum [4], and Oracle Exadata [6] has analysis application.
- **In-Memory Systems:** SAP HANA [29] can provide in memory computing which is performance oriented data processing method.
- **Columnar:** SAP's Sybase IQ Hewlett Packard's Vertica, Paraccel, Infobright, Exasol, Calpont, and Sand [24] have the data stored by column rather than row manner that access to information easier.

- Hadoop Distributions: Hadoop Distributed File System (HDFS) [15] store data across commodity machine cluster, while providing high availability and fault tolerance using distributed file system.

- NoSQL Databases: These database mean not only SQL servers.
  - **Key Value Pair Databases:** Cassandra, Hbase, and Basho Riak [24] are this type of databases.
  - **Document Stores:** JSON, MongoDB and Couchbase [24] belong to this type of databases.
  - **SQL MapReduce:** Teradata's Aster Data [28] and EMC Greenplum [4] have data processing features using MapReduce.
  - **Graph Systems:** These systems contact people through multimedia companies.
  - **Unified Information Access:** Attivio, MarkLogic, and Splunk can handle both structured data and unstructured data.
  - **Other:** There are also many NoSQL databases resulting from the different uses of application and information [24].

3.2 Infrastructure and Analytics

Infrastructure and analytics can be distinguished in following layers.

- Data Storage Layer: Data storage layer [30] can be described as follows.
  - **Hbase:** Hbase [4] is an open source distributed database that Google run large data model written in Java [31, 32].
  - **S3:** Amazon S3 (Simple Storage Service) is provided by online file storage Web Services such as REST，SOAP and Bit Torrent [33].
  - **HDFS:** Hadoop Distributed File System is inspired by the Google File System which can store very large files in the machines across a large cluster [34].
  - **Tachyon:**Tachyon [21] is a memory centric distributed storage system to achieve reliable data stored in the cache shared across a cluster framework [22]. Tachyon avoids frequent disk read load data sets. And is also compatible with Hadoop. File system content is stored in memory of the body of all cluster nodes. Thus, the system achieves higher storage speed than that of traditional disk -based system like HDFS [32].

- Data Processing Layer: Characteristics for data processing layer can be found as follows.
  - **Yarn:** MapReduce after Hadoop 0.23 has a comprehensive overhaul of MapReduce which become MRv2 or Yarn. The basic concept is to give MRv2 JobTracker, resource management and job scheduling / monitoring two functions into separate programs [4, 16]. It has ResourceManager (RM) and ApplicationMaster (AM). The two main components are the Scheduler and Applications Manager [35].
  - **Mesos:** Apache Mesos is an open source cluster manager developed at the University of California Berkeley. It provides an effective isolation and sharing of resources across a distributed application or

framework [4, 36]. Currently, there are at least 50 organizations use Mesos including some large companies Twitter, Airbnb and Apple [37].

- **Map Reduce:** MapReduce is a programming model which is for processing and generating large data sets related to the implementation of parallel, distributed algorithm across the cluster [2, 4, 38]. MapReduce programs executed by a Map ( ) method of filtering and sorting and Reduce ( ) method, the executive summary of the operation. Map Reduce is a part of Apache Hadoop [38].
- Yarn and Mesos pros and cons [39] can be seen as in below table.

Table 2 Yarn and Mesos

		Yarn	Mesos
1	MapReduce [4, 39]	java	C++
2	Data processing [39]	In memory	In memory + CPU
3	Operating system [39]	Unix	Linux
4	Setups [40]	Specific instructions	Flexible
5	Coding [40]	3 times	1 time
6	Cluster [4, 40]	Hadoop like	Specific infrastructure
7	Security [4, 40]	Kerberos basic inheritance Hadoop security	Lack of implementation
8	Deployment [40]	Direct	Cannot directly deployed
9	Documentation and specs [40]	Less	Mature
10	Usage on Hadoop [40]	Transparent	Less transparent
11	Usability [40]	Yahoo, Hortonworks	Self-applications which have compatibility concerns.

- Data Access Layer: Characteristics for data access layer can be found as follows.
  - **Sqoop:** Sqoop acts in between relational databases and Hadoop command -line interface for transmitting data applications [41]. It supports a single table or SQL query incremental load and can run into the database which is updated several times to save jobs. Import can be used to fill Hive or table of Hbase [4, 41]. Export to Hadoop data can be converted into a relational database. Microsoft uses Sqoop based connectors to help transport from Microsoft SQL Server databases to Hadoop. Couchbase has provided Couchbase Hadoop server connector with Sqoop device.
  - **Hive:** Apache Hive [4, 42] is establishing a database on top of Hadoop infrastructure to provide data collection, query and analysis [32]. Apache Hive which has been first developed by Hive has now been used by other companies like Netflix and other development [43]. Amazon is also using Apache Hive software on Amazon Web Services for Amazon Elastic Map Reduce.
  - **Pig:** Apache Pig [44] is the data analysis program which consists of a high-level language for analyzing data, infrastructure and the platform used to evaluate those programs with large data sets [45]. Their structure is suitable for a large number of parallel processing, thus enabling them to handle very large data sets [4, 32]. Main features include easy to program and optimize the opportunities and scalability.
  - **Hbase:** Hbase [4] is an open source distributed database that Google run large data model written in Java [31, 32].
  - **Storm:** Apache Storm [46, 47] a free open source distributed real-time computing systems. Storm has a lot of use cases such as realtime analytics, online machine learning,

continuous computing, distributed RPC, ETL, and more[4, 32].

- **Avro:** Apache Avro [48, 4] is a data serialization system. Avro provides:
  1.Rich data structures.
  2.Small and fast binary data format.
  3.Container type files to store persistent data.
  4.Remote procedure call（RPC）.
  5.Simple dynamic language integration.
- **Map Reduce:** MapReduce is a programming model for processing and generating large data sets related to the implementation of parallel , distributed across the cluster algorithm [2, 4, 38]. MapReduce is executed by a Map( ) method of filtering and sorting and Reduce( ) method, the executive summary of the operation. Map Reduce is part of the Apache Hadoop [38].
- **Spark:** Apache Spark [4, 49, 50] a fast and large-scale data processing engine. It was originally developed at the University of California Berkeley campus AMPLab and is an open source cluster computing framework. It was donated to the Apache Software Foundation. Comparing to two-stage disk of Hadoop MapReduce, the in-memory operation of Spark for certain applications provide performance up to 100 times faster.

- Data Management Layer: Characteristics of data management layer can be found as follows.
  - **Oozie:** Oozie [4, 27, 51] is Direct Acyclical Graphs(DAG) of MR workflow scheduling system . Its coordinator can offer job by time (frequency) and data availability.
  - **Chukwa:** Chukwa [27, 52, 53] is a large-scale log aggregation and analysis system. It is also an open source data collection system for monitoring large-scale distributed systems. Chukwa is built on Hadoop Distributed File System (HDFS) and Map Reduce framework and thus also inherits the scalability and robustness of Hadoop. It includes a flexible and powerful toolkit for displaying, monitoring and analyzing of the results, so that the data collected can be fully utilized.
  - **Flume:** Flume [4, 27, 54] a distributed, reliable service which can be used effectively to collect and aggregate a large number of mobile service log data.
  - **Zookeeper:** Zookeeper [4, 27, 55] is used to maintain configuration information, naming, providing distributed synchronization, and services for centralized services. All of these types of services use distributed programs or being used by another application.
  - **Spark SQL:** Spark SQL [4, 56] replaced Shark SQL. Spark SQL query can plan structured data, or use the familiar SQL data frames API available in Java, Scala, Python and R.
  - **Mlib:** MLlib [57] is suitable for Spark API using Python, NumPy and starts at Spark 0.9 interoperability .Hadoop can use any data source (such as HDFS, HBase or local files). It is easy to insert into Hadoop workflows. And its algorithm is a hundred times faster than that of Map Reduce.
  - **Logstash:** Logstash [4, 58] is an event management and logging tool. It is generally used in a larger system log collection, processing, storage and search activities.
  - **Scribe:** Scribe [59, 60] is a server used to gather real-time streaming data from a large number of server log data. It can be modified without the client's scalable, and robust fault or any particular machine on the network.

- Data Analysis Layer: Characteristics of data analysis layer can be found as follows.
  - **EDW:** Enterprise data warehouse (EDW) [61] is a system for reporting and data analysis. It is the central repository from one or more different sources of comprehensive data. They store current and historical data, and is used across the enterprise to create a knowledge-based analysis. Examples may include reports from comparisons of trends from quarterly and annually data to detailed daily sales analysis.
  - **Mahout:** Apache Mahout's [27, 62] goal is to build high-performance machines to quickly create scalable applications learning environment. The three main components are scalable algorithms, new Scala + Spark and H2O (Apache Flink ongoing) algorithm and Mahout which is a mature Hadoop's MapReduce algorithm.
  - **Hive:** Apache Hive [4, 42] is to establish a database on top of Hadoop infrastructure to provide data collection , query and analysis [32]. Apache Hive which is initially developed by Facebook has now been used by other companies like Netflix and other development [43]. Amazon is also using Apache Hive software on Amazon Web Services for Amazon Elastic Map Reduce.
  - **Impala:** Impala [4, 63] is Cloudera's open source on Apache Hadoop cluster of computers which run massively parallel data processing (MPP) of the SQL query engine .
  - **Stinger:** Stinger [4, 64] is a continuation of speed, scale and a familiar SQL in the Hive. The three delivery schedules can be achieved in an open community of Apache Hive. These three goals are speed, size and SQL query.
  - **Drill:** Apache Drill [4, 27, 65] supports data-intensive distributed applications interactive analysis of large data sets of open-source software framework. It is an open source version of Google Dremel system and it can be used as Google Big Query of infrastructure services.

- Data Visualization Layer: Data visualization is an important layer of big data. Comparison [66] on this can be found as follows.
  - **SAS:** Statistical Analysis System (SAS) [67, 68] is a high level analysis system developed by the SAS Institute to use in software development, multivariate analysis, business intelligence, data management and predictive analysis software packages.
  - **SAP:** Systems, Applications & Products in Data Processing (SAP) [29, 69] is an application server that includes an in memory, column-oriented, relational database management system. It is previous known as "SAP High-Performance Analytic Appliance".
  - **Qlikview:** Qlikview [70] is an in-memory, business discovery tool which also is a Business Intelligence application that helps organizations big and small in data discovery. QlikView can deliver data visually which provides context of the data in rich but simple format. It consists of QlikView desktop, server and publisher.
  - **Tableau:** The visualization capabilities of Tableau [71, 72] are diverse and highly insightful. Tableau features such as "word clouds" or "bubble maps" are great to enhance comprehension. Its tree maps also provide the facility to add context for graphics. The tree

maps are mainly used to display relative proportions of multiple categories of a variety of information. There also is a capability for laying out the dashboard via “overlaps” is also a big powerful feature. It enables efficient use of screen space.

- **D3.js:** D3.js(Data-Driven Documents) [73, 74] can produce dynamic, interactive data visualization using JavaScript library on Web browser. It uses a widely implemented SVG, HTML5 and CSS standards. Even from its earlier Protovis framework. D3.js can control the final visual effect.
- **Elk:** Combining popular Elasticsearch, Logstash and Kibana, Elasticsearch company has established an end - to-end ELK stack [75, 76]. It provide immediate actionable insights for any type of structured and unstructured data sources.
- **Datawatch:** Datawatch [77] has the technique which relies on in-memory OLAP (Online Analytical Processing) perspective, which includes a tree display through a series of visualization. This allows the user to load data, select variables and hierarchy, and navigate through the resulting visualization, filtering, amplification and drilling (also known as slicing and cutting), to identify outliers, correlations and trends.

- Business intelligence: Business intelligence platforms [78, 79] are also very important in big data and benchmarking can be found as follows.
  - **Pentaho:** Pentaho [80, 81] provides business analysis. It is an open source business intelligence (BI) product that provides data integration, OLAP services, reporting, monitoring, data mining and ETL functionality. Hitachi acquired Pentaho in 2015.
  - **SAP:** Systems, Applications & Products in Data Processing (SAP) [29, 69] is an application server that includes an in memory, column-oriented, relational database management system. It is previous known as "SAP High-Performance Analytic Appliance". SAP attracts larger organizations with complicated needs.
  - **Cognos:** Cognos [82, 83] can resolve to help understand , monitor and manage business performance including business reporting and analysis, profitability measurement, budgeting, forecasting and optimization of cost management. It is a fast and effective technology to provide multi-dimensional business intelligence data.

4 CASE STUDY- FACEBOOK

Facebook [11, 84] is collecting data from two sources. MySQL tier contains user data whereas web servers generate event based log data. Web server data is collected to Scribe servers and then executed in Hadoop clusters. The aggregated log data from Scribe server is written to HDFS. The data in HDFS is compressed periodically and transferred to Hive Hadoop for further processing. Data analysis queries in Facebook are specified with graphical user interface called HiPal or Hive command line interface.

Table 3 Facebook Infrastructure

	Architecture
B.I. [11, 85]	Pentaho, SAP Cognos
Visualization [11, 85]	EDW, SAS, SAP, Qlikview, Tableau, D3.js, Kibana, Datawatch, Project Palantir, Friend Wheel, Touch Graph Browser, Mutual Friend network, Nexus, HiPal
Analysis [11, 84, 85]	EDW Connector, EDW, Hive, Impala, Stinger, Drill Mahout, Mlib,
Data Mgmt. [11, 85]	Oozie, Chukwa, Flume, Flumetd, Zookeeper, Spark SQL, Scribe, Logstash, Elasticsearch, HiPal, Databee

Data Access[11, 86, 87]	Sqoop, Hive, Pig, Hbase, Apache Storm, Avro, Presto, Giraph
Data Processing[11, 40]	Hadoop Map Reduced, Hadoop Yarn, Cassandra
Data Storage[11, 87]	HDFS, S3, Hbase, Scuba

Table 4 Facebook Infrastructure Case Study

B.I.	Pentaho[1]	SAP Business[1]	Cognos[1]							
Visualization	EDW[1]	SAS[1]	SAP[1]	Qlikview[1]	Tableau[1]	D3.js[1]	Kibana[1]	Datawatch[1]		Project Palantir, Friend Wheel, Touch Graph Browser, Mutual Friend network, Nexus, HiPal
Analysis	EDW Connector[1]	EDW[1]	Hive[1]	Impala[1]	Stinger[1]	Drill[1]	Mlib[1]	Mahout[1]		
Data Mgmt	Oozie[1]	Chukwa[1]	Flume[1]	Flumetd[1]	Zookeeper[1]	Spark SQL[1]	Elastic search[1]	Logstash[1]	Scribe[1]	HiPal, Databee
Data Access	Sqoop[1]	Hive[1]	Pig[1]	Hbase[1]	Apache Storm[1]	Avro[1]	Hadoop Map Reduced[2]	Spark[2]		Presto, Giraph
Data Processing	Hadoop Map Reduced[1]	Hadoop Yarn[1]	Mesos[2]							Cassandra
Data Storage	HDFS[1, 2]	S3[1, 2]	Hbase[1, 2]	Tachyon[2]						Scuba

Notes: 1: Open Source Hadoop Stack 2. BDAS (Berkeley data analytics stack) Right most column is the proprietary components.

5 CONCLUSION

From the case study above, it is found that most of the real world application can be based on our general big data framework such as HDFS in storage layer, HIVE and PIG in data access layer, Tableau and Qlikview in visualization layer but other components maybe customized into proprietary software or by suggestions of domain experts.

The overall framework hierarchy and characteristics discussed with experts remain the same. Only difference may be specific needs and reference architecture of different proprietary domain. For example, even the same type of social media big data companies such as Facebook and Twitter has very similar general framework but they have separate proprietary customized software.

ACKNOWLEDGE

The authors gratefully acknowledge the support from Taiwan Information Security Center (TWISC) and Ministry of Science and Technology (MOST), under the grants MOST , 105-2923-E-182 -001 -MY3, 105-2221-E-182-053, 105-2632-H-182-001, and 105-2221-E-034 -020 -.

REFERENCES

[1] S. Ghemawat, H. Gobioff, and S.-T. Leung, “The Google File System,” in Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles. 2003: New York, NY, USA.

[2] J. Dean and S. Ghemawat, “MapReduce: A Flexible Data Processing Tool,” Communications of the ACM, vol. 53, no. 1, pp. 72-77, 2010.

[3] Apache Software Foundation, Hadoop 1.2.1 Documentation, Available from: http://hadoop.apache.org/docs/r1.2.1/index.html.

[4] T. White, Hadoop: The Definitive Guide. the 4th Edition. 2015: O' Reilly Media.

[5] IBM website, IBM Stream Computing, Available from: https://www.ibm.com/analytics/us/en/technology/stream-computing/.

[6] Oracle website, Oracle and Big Data: Big Data for the Enterprise, Available from: https://www.oracle.com/ big-data/index.html

[7] Hewlett Packard Enterprise website, Big Data Solutions, Available from: https://www.hpe.com/us/en/solutions/big-data.html.

[8] Miscrosoft website, Big Data, Available from: http://www.microsoft.com/enterprise/it-trends/big-data/default.aspx#fbid=m1BCtxTTPjr.

[9] A. Tattersall and M.J. Grant, "Big Data - What Is It and Why It Matters," Health Information & Libraries Journal, vol 33, no. 2, pp. 89-91, 2016.

[10] Cloudera website, Cloudera Enterprise: The World'S Most Popular Apache Hadoop Solution, Available from: http://www.cloudera.com/content/www/en-us/products.html.

[11] P. Pääkkönen and D. Pakkala, "Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems," Big Data Research, vol. 2, no. 2, pp. 166-186, 2015.

[12] D. Singh and C.K. Reddy, "A Survey on Platforms for Big Data Analytics," Journal of Big Data, vol. 1, no. 8, 2014.

[13] Gartner, Inc., Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data, Available from: http://www.gartner.com/newsroom/ id/1731916.

[14] Y. Demchenko, C. de Laat, and P. Membrey, "Defining architecture Components of the Big Data Ecosystem.," in Proceedings of 2014 International Conference on Collaboration Technologies and Systems (CTS), 2014, pp. 104-112.

[15] Apache Software Foundation, HDFS architecture guide, Available from https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html#Introduction.

[16] V.K. Vavilapalli, A.C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, B. Saha, C. Curino, O. O'Malley, B. Reed, and E. Baldeschwieler, "Apache Hadoop YARN: Yet Another Resource Negotiator," in Proceedings of the 4th annual Symposium on Cloud Computing, 2013, Article No. 5.

[17] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Communications of the ACM, vol. 51, no. 1, pp. 107-113, 2008.

[18] K.H. Lee, Y.J. Lee, H. Choi, Y.D. Chung, and B. Moon, "Parallel Data Processing with MapReduce: A Survey," ACM SIGMOD Record, vol. 40, no. 4, pp. 11-20, 2011.

[19] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, "Pig Latin: a Not-so-foreign Language for Data Processing," in Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, 2008, pp 1099-1110.

[20] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy, "Hive: a Warehousing Solution over a Map-reduce Framework," in Proceedings of the VLDB Endowment, vol. 2, no. 2, 2009, pp. 1626-1629.

[21] AMPlab UC Berkeley, Berkeley Data Analysis Stack, Available from: https://amplab.cs.berkeley.edu/software/.

[22] Alluxio website, Open Source Memory Speed Virtual Distributed Storage, Available from: http://http://www.alluxio.org/.

[23] Apache Software Foundation, Aurora Project Incubation Status, Available from https://incubator.apache.org/projects/aurora.html.

[24] W.Eckerson, Categorizing Big Data Processing Systems, Beye Network, Avaiable from: http://www.-eye-network.com/blogs/eckerson/archives/2012/02 categorizing_bi.php

[25] M.B. Nirmala, "A Survey of Big Data Analytics Systems: Appliances, Platforms, and Frameworks," Handbook of Research for Cloud Infrastructures to Big Data Analytics, IGI Global, 2014, p.p. 393-419.

[26] IBM website, Big Data- Big Data Solutions That Work for You, Available from: http://www-03.ibm.com/software/products/en/category/bigdata.

[27] M. Maier, Towards a Big Data Reference Architecture, Master's thesis, Eindhoven University of Technology, 2013.

[28] W. O'Connell, I.T. Ieong, D. Schrater, C. Watson, G. Au, A. Biliris, S. Choo, P. Colin, G. Linderman, E. Panagos, J. Wang, and T.Walter, "A Teradata Content-based Multimedia Object Manager for Massively Parallel Architectures," in Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, 1996, pp. 68-78.

[29] SAP website, SAP HANA, Available from: http://go.sap.com/index.html.

[30] S. Neumann, Storing Apache Hadoop Data on the Cloud - HDFS vs. S3, Available from: https://www.xplenty.com/blog/2014/03/storing-apache-hadoop-data-cloud-hdfs-vs-s3/.

[31] Apache Software Foundation, Apache HBase, Available from: http://hbase.apache.org/.

[32] J. Roman, The Hadoop Ecosystem Table, Available from https://hadoopecosystemtable.github.io/.

[33] Amazon, Amazon S3, Available from: http://aws.amazon.com/tw/s3/.

[34] Hadoop Wiki, Hadoop Distributed File System, Available from: https://wiki.apache.org/hadoop/HDFS.

[35] Apache Software Foundation, Apache Hadoop YARN, Available from: https://hadoop. apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html.

[36] Wikipedia, Apache Mesos, Available from: https://en.wikipedia.org/wiki/Apache_Mesos.

[37] Apache Mesos, Organizations Using Mesos, Available from http://mesos.apache.org/documentation/latest/powered-by-mesos/

[38] Wikipedia, Map Reduce, Available from: https://en.wikipedia.org/wiki/MapReduce.

[39] Quora, How Does YARN Compare to Mesos?, Available from: https://www.quora.com/How-does-YARN-compare-to-Mesos.

[40] Cloudtimes, Facebook's Big Data: New Concept in Data Management, Available from: http://cloudtimes.org/2012/09/03/facebooks-big-data-new-concept-in-data-management/.

[41] Apache Software Foundation, Apache Sqoop, Available from: http://sqoop.apache.org/.

[42] Apache Software Foundation, Apache Hive, Available from: https://hive.apache.org/.

[43] Wikipedia, Hive, Available from: https://en.wikipedia.org/wiki/Apache_Hive.

[44] Apache Software Foundation, Apache Pig, Available from: https://pig.apache.org/.

[45] Wikipedia, Pig (Programming Tool), Available from: https://en.wikipedia.org/wiki/Pig_(programming_tool).

[46] Apache Software Foundation, Apache Storm, Available from: http://storm.apache.org/.

[47] Wikipedia, Storm (event processor), Available from: https://en.wikipedia.org/wiki/Storm_(event_processor).

[48] Apache Software Foundation, Apache Avro, Available from: https://avro.apache.org/docs/current/.

[49] Apache Software Foundation, Apache Spark, Available from: http://spark.apache.org/.

[50] Wikipedia, Spark, Available from: https://en.wikipedia.org/wiki/Apache_Spark.

[51] Apache Software Foundation, Apache Oozie Workflow Scheduler for Hadoop, Available from: http://oozie.apache.org/.

[52] Apache Software Foundation, Apache Chukwa, Available from: https://chukwa.apache.org/.

[53] J. Boulon, A. Konwinski, R. Qi, A. Rabkin, E. Yang, and M. Yan, "Chukwa: a large-scale monitoring system," in Proceedings of Cloud Computing and its Applications (CCA '08), 2008.

[54] Apache Software Foundation, Apache Flume, Available from: https://flume.apache.org/.

[55] Apache Software Foundation, Apache Zookeeper, Available from: https://zookeeper.apache.org/.

[56] Apache Software Foundation, Spark SQL, Available from: https://spark.apache.org/sql/.

[57] Apache Software Foundation, Mlib, Available from: http://spark.apache.org/mllib/.

[58] Elastic, Logstash, Collect, Enrich & Transport , Available from: https://www.elastic.co/products/logstash.

[59] Quora, What's the Difference between Thrift and Scribe?, Available from: https://www.quora.com/Whats-the-difference-between-Thrift- and-Scribe.

[60] Wikipedia, Scribe (log server), Available from: https://en.wikipedia.org/wiki/ Scribe_(log_server).

[61] Wikipedia, Data Warehouse, Available from: https://en.wikipedia.org/wiki/Data_warehouse.

[62] Apache Software Foundation, Apache Mahout, Available from: http://mahout.apache.org/.

[63] Cloudera. Apache Impala (Incubating), Available from: http://www.cloudera.com/content/www/en-us/products/apache-hadoop/impala.html.

[64] Hortonworks. Enterprise Big Data Solutions, Available from: http://hortonworks.com/innovation/stinger/.

[65] Apache Software Foundation, Apache Drill, Available from: https://drill.apache.org/.

[66] E.M. Forster, G. Wallas, and A. Gide, Data Visualization- Discover, Analyze, Explore, Pivot, Drilldown, Visualize Your Data…"How do I know what I think until I see what I say?", Available from: https://apandre.wordpress.com/tools/comparison/.

[67] SAS website, Analytics Software & Solutions, Available from: http://www.sas.com/zh_tw/home.html.

[68] Wikipedia, SAS(Software), Available from: https://en.wikipedia.org/wiki/SAS_(software).

[69] Wikipedia, SAP SE, Available from: https://en.wikipedia.org/wiki/SAP_SE.

[70] Qlik websiet, Qli, Available from: http://www.qlik.com/.

[71] Tableau website, Tableau, Available from: http://www.tableau.com/.

[72] Wikipedia, Tableau Software, Available from: https://en.wikipedia.org/ wiki/Tableau_Software.

[73] Wikipedia, D3.js, Available from: https://en.wikipedia.org/wiki/D3.js.

[74] Mike Bostock, Data-Driven Documents, Available from: http://d3js.org/.

[75] Christopher'blog, Visualizing Data With Elasticsearch, Logstash and Kibana, Available from: https://blog.webkid.io/visualize-datasets-with-elk/.

[76] Scaleway, How to Collect and Visualize Your Log with ELK Stack., Available from: https://www.scaleway.com/docs/how-to-use-the-elk-stack-instant-apps/.

[77] DataWatch website, DataWatch, Available from: http://www. datawatch.com/.

[78] Business Application Research Center, The BI Survey 10, p.p. 4-19, 2011.

[79] Butler Analytics, Enterprise BI Platforms Compared, Available from: http://www.butleranalytics.com/enterprise-bi-platforms-compared/.

[80] Pentaho, A Comprehensive Data Integration and Business Analytics Platform, Available from: http://www.pentaho.com/.

[81] Wikipedia, Pentaho Suite, Available from: https://en.wikipedia.org/wiki/Pentaho.

[82] GoliInfo, Cognos Business Intelligence and Enterprise Performance Management, Available from: http://www.cognos-bi.info/.

[83] IBM, Cognos, Available from: http://www-01.ibm.com/software/analytics/cognos/.

[84] A. Thusoo Z. Shao, S. Anthony, D. Borthakur, N. Jain, J.S. Sarma, R. Murthy, and H. Liu, "Data Warehousing and Analytics Infrastructure at Facebook," in Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, 2010, p.p. 1013-1020 .

[85] S. Schroeder, 6 Gorgeous Facebook Visualizations, Available from: http://mashable.com/2009/08/21/gorgeous-facebook-visualizations/#Rg9f0Wo7FOqG.

[86] L. Chan, Presto: Interacting with Petabytes of Data at Facebook, Available from: https://www.facebook.com/notes/facebook-engineering/presto-interacting-with-petabytes-of-data-at-facebook/10151786197628920/.

[87] J. Wiener and N. Bronson, Facebook's Top Open Data Problems, Available from: https://research.facebook.com/blog/facebook-s-top-open-data-problems/.

Thinning Round-robin with Rating Index and Virtual Environment for Battle in Applied Java Programming Exercise with Game Strategy and Contest Style

Naoki Hanakawa and Hiroyuki Tominaga
Kagawa University
2217-20, Hayashi-machi, Takamatsu City, Kagawa Pref., Japan
s16g464@stu.kagawa-u.ac.jp, tominaga@eng.kagawa-u.ac.jp

ABSTRACT

We have proposed an applied Java programming exercise with board-game strategy for problem solving learning. During implementation of hand methods of Gogo game, students learn realization of ideas as algorithms and revision with trial and error by execution results. We have developed support system WinG. WinG-LA is a local review tool. It offers a game execution library as Java API and contains four modules for examination of a strategy. WinG-CS is a contest management server. It executes a lot of games among uploaded students' programs. It maintains a preliminary and the final period for battle league. They decide students' scores by the result of round-robin matching. We performed an educational practice in 2011. In this paper, we describe about improvement of WinG-CS and exercise practices. We introduce virtual environment into server for efficiency. We also reconfigure database.

KEYWORDS

Java programming exercise, board game strategy, thinning round-robin in contest style, rating index, continuous integration

1 INTRODUCTION

1.1 The final goal of programming education

In an information engineering college, a programming exercise is regarded as a required subject. After an introductory exercise in a junior class, an advanced exercise in a senior class treats object-oriented programming with C++/Java language.
We consider that two aspects as process and product are important in programming education for the final goal. One is the algorithmic approach. It means that programming skill is regarded as an effective tool for problem solving method. Another is the desirable attitude as an engineer. It means continuous integration of programming according to software development method.
However, if subjects of the field are so far from student's interest, it is difficult to illustrate concrete images and to pursue the goal of programming. Moreover, a student does not feel attachment to a software product which he creates and has to maintain.
A teacher often picks up applied problems related with some specific IT fields. He also treats simple projects about information system based on actual works. However, a student often does not understand mathematical concepts as required knowledge. Or, he does not feel concrete image because of lacking social experience. Therefore, he tends to stumble on a clue of the problem condition and gives up solving by programming.

1.2 Programming exercise of game strategy with contest style

In the field of knowledge information processing, game strategy programming is focused as an attractive exercise subject. It includes several learning items, such as, formulation of move operation, condition of application, pattern matching of board situation, intellectual searching by forecast, and simulation trial.
It includes the competitive learning approach. It adopts a contest style of strategies among students each other. A student feels them voluntary attempts rather than teacher's imperative tasks. The approach is expected to raise motivation and to bring educational effects. Moreover, the best strategy as the very

correct answer is not clear. So, learners are required continuous integration by competitions with opponents.

2 THE OUTINE OF OUR EXERCISE

2.1 Applied Java exercise of board game strategy

We have proposed an applied programming exercise of board game strategy with a contest style. We adopt board game "Gogo", which is a variant of "Ninuki-Renju" like "Pente". We have developed support system WinG [1][2]. We have carried out several educational practices since 2005 in an applied programming lesson. This exercise became a required subject, and it was positioned as a more serious problem since 2011.

2.2 Rules and features of board game Gogo

Japanese traditional gobang (Gomoku-Narabe) is a very famous board game, which is a simple version of "Renju". A player puts a stone with his color (white or black) in a square mesh board alternately. He aims to create a connected run with his five stones in a straight or diagonal line. "Ninuki-Renju" is not so popular but a very interesting game based on Gomoku. It has added rules about removing and capturing pinched two opponent's stones. "Pente" is known as a variant of the game in Germany and Poland [3].
We adopt "Gogo", which is also a simple version of Ninuki-Renju. The board has 169 cells in 13 square sizes. We adopt some arrangement of the rules for an affordable subject of programming exercise (Fig.1). The winning condition has two ways, creating "steady" five-run or capturing five pairs (ten stones). As an "unsteady" five-run has removable pairs, you can break the run in the next turn and the game does not finish. Moreover, if you obtain the fifth pairs in the situation, you win as come-from-behind victory. Over five runs are not applied in the winning condition. Making double three-runs (San-San) by putting a stone is a mismove as the losing condition, while the one by removing a pair is allowable.
Gogo is so profound that removing stones cause significant change of board situation and surprising turnover of a game aspect. When you think of some ideas of a strategy, you must consider two aspects. One is a tendency to aim connecting or capturing, and the other is a tendency to choose attack or defense. The policy of strategy has a lot of variations with various evaluation functions. Though the rules are simple, a beginner shows his individuality with the preference and attitude. As documents about rotes of Ninuki-Renju are not enough, you must search good strategies by yourself.

2.3 The previous system and our research

We developed the first version WinG in 2005. It consisted of local execution environment and contest management server WinG-CS [1]. The environment offers game execution between strategies or human player, and review functions of battle record. By doing this, we support students to create strategies. With the initial edition of WinG-CS, we devised a mechanism that students submit strategies and execute on the server.
The environment changes a local development tool WinG-LA as the next version [2]. WinG-LA consists of 4 modules, such as, "game execution", "battle record replay ", "board situation generation", and "hand trial test". We introduced SWG and the first version of WWG (weighted winning grade), which is a simple rating index [4].

2.4 Making progress of strategy program

We offer a board game execution library for Gogo as Java API. He overrides hand method calc_hand() in his subclass inherited Computer Player class. The method receives a current state as an argument. It returns a next hand. The state as an instance of State class consists of board situation for stone placement and two pockets for captured stones. An instance of Master class manages a game progress according to the rules. We present prototype of strategy to student, which includes necessary

processing as comment. Almost strategy codes have 300 to 500 lines.
The overview of strategy design in Gogo is shown in Fig. 2. In the first step, you consider the outline of a strategy idea and decide a tactics policy. Each tactic is almost described by if-then rules in knowledge information processing. The left-hand side of a rule as the second step is pattern recognition of stone placement on a board. You realize various matching algorithms of stone placement such as a four-run, double runs and multiple capturing. You may refine more detailed patterns and find specific patterns for a winning process. The right-hand side as the third step is an assignment of an evaluation value for every cell by a heuristics function. The evaluation value must be revised by trial-and-error by game execution. In the fourth step, you consider the global board situation and an adjustment of the priority of the rules. You may adopt a look-ahead searching algorithm and probability approach. In the last step, a cell with the maximum value is selected as the hand.

2.5 A preliminary and the final league in game contest

In our exercise, we have set a tournament period of several weeks. We also carry out preliminary league and final league in battle style. Learners can submit their strategy many times during about 5 weeks in preliminary league. Submitted strategy battles other strategies on contest management server. Battle result is given by weight winning grade considering rating of each strategies. In preliminary league, the server receives about 1000 submissions in 40 learners. It means the server performs hundreds of thousands of battles in round-robin battle. It needs a large of execution time. So, we introduce thinning round-robin battle in consideration of winning degree.
Contest management server WinG-CS publishes all battle results and ranking of submitted strategies. Learners can check steps and cause of victory in battle results and replay battle records excluding opponent strategy codes. We introduce index strategies in 3 strength levels as strength index. These information is effective to feedback for revising strategies. In this way, we motivate learners to revise program continuously by providing the opportunity to always evaluate their own strategies.
After finished preliminary league, students comprehensively judge, and select their own best strategy which participate in final league. In final league, the server performs round-robin battle with the strategies. the server also let result reflect. An evaluation of each student is decided by the score and a summary report, in which he/she analyzes the process and the result.

3 Purpose and position of our exercise

3.1 Game strategy programming exercises in our curriculum

We have carried out this exercise in experiment for the third-grade students in an information and electronics engineering college. Students already studied data structures and algorithms in C, and basic grammar of Java in other class. And also, we have carried out an exercise with card game strategy in C [5]. Its subject matter is poker game. Students aim for high points, and they are ranked by points. Students learn a design of data structures such as deck and hand, implementation of pattern-match based on production rule, and judgement of poker hands. After finished poker exercise, we start this exercise.
We describe differences between card game exercise and board game exercise. these exercises complement each other. In card game, we use incomplete information game. In addition, card game exercise is a score style. So, students' effort tends to directly lead to results. On the other hand, in this exercise, we use perfect information game. And, this exercise is a battle style. So, it is difficult to improve result by their own efforts.

3.2 Learning items and educational purpose

We describe our educational goals of this exercise in two aspects: analysis and design, implementation and verification. On analysis and design, we make students examine application condition of hand and priority of

processing based on search algorithms in knowledge engineering. For example, the to formulate of application condition of hand, to extract pattern of characteristic state participating in win or lose such as San-San, to implement heuristic knowledge obtained from actual experience.
On implementation and verification, we make students to practice spiral development method based on object-oriented programming. In other words, we stimulate students to improvement evaluation of state or hand from battle result. For example, to localize points to fix evaluation parameters using modularization by method, to verify the validity of method of hand using sample state as test case, to improve by own state of win or lose or replaying battle records.

3.3 Related works

We describe related competitions. IPSJ have held SamurAI Coding which is game AI programming contest since 2012 [6]. IEEE CIG have held many game AI competitions every year [7]. MIT also have held Battlecode which is battle game AI contest [8]. In general corporations, CodinGame has held CodinGame Contest [9]. It is an open competition. So, the purpose is enhancement of user's programming skill through creating game AI. Thus, many strategy programming contests are held in the world.
We also describe related research about applications to education. Canada aims for improvement of knowledge of students about AI trough international online contest [10]. In this paper, they verified the educational efforts about participating Google AI Challenge as part of class. It realized about improving motivation and result by using contests in the class about AI. Yoon teaches basic concepts about game and AI using Angry Birds [11]. Other researchers practice classes using computer game with contest style [12][13]. In our research, the purpose is not only creating strong strategy but also learning rightful techniques or good manners about software development. This is the most different point.

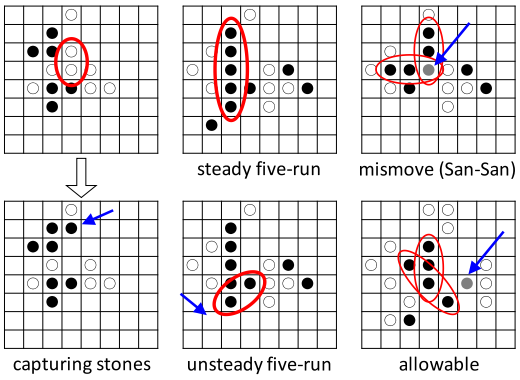


Figure 1. The rules of Gogo and board state

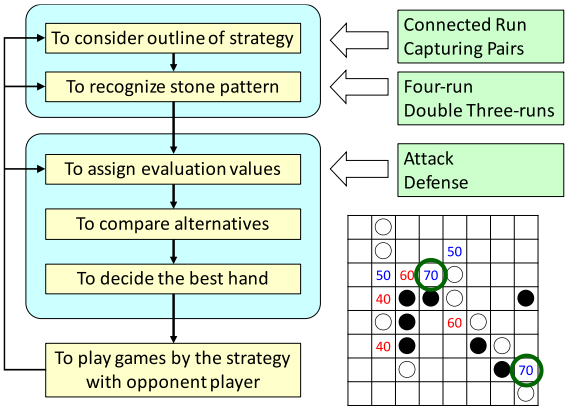


Figure 2. The overview of strategy design in Gogo

4 A SUMMARY OF CONTEST MANAGEMENT SERVER WinG-CS

4.1 Support system WinG and execution environment

To realize the exercise, we have developed support environment WinG(Fig.3). Learners download WinG-LA, which is support for developing strategies, from shared folder on our server. They implement their strategies on their own PC. We offer a game execution library as Java API, their development environment is freedom. WinG-LA provides some modules such as battle execution, debugging efficiently. It also provides some samples strategies and board situations for their strategies examination.
Server side WinG-CS, which support for contest management, manages some processes on server side in preliminary leagues and final leagues. League corresponds each exercise in

ISBN: 978-1-941968-42-0 ©2017 SDIWC

each year. The system makes submitted strategies meet them in battle game, and it publishes battle results and records (Fig.4). The system saves battle results battle records to database on the server. Learners can browse these data. They also can download battle records, and replay on WinG-LA.

4.2 Functions and configuration of WinG-CS

In this paper, we describe features and enhancement in 2016 about WinG-CS. During from 2011 to 2015, we had carried out educational practices with previous version of WinG-CS. The system was implemented in Ruby 1.9, and the database is saved by XML. New WinG-CS, which is enhanced in 2016, is implemented in Ruby 2.3 and Ruby on Rails 4.2. Rails is based on Model-View-Controller architecture, so we can reduce development cost in coordination with database and GUI. our contest server mounts Intel Core i7 as CPU, and DDR3 32GB as RAM. OS of the server is Linux series CentOS 6.8, but we update the Linux kernel to 3.10 to use virtual environment Docker.

4.3 Internal processing of WinG-CS

WinG-CS accepts source codes of learners' strategy during preliminary league. WinG-CS processes acceptance processing and battle processing as internal processing at all such times. acceptance processing and battle processing are proceeded seamlessly after submission.

Acceptance processing performs static check of synchronous processing and dynamic check of asynchronous processing. WinG-CS registers the submission file to submission DB, and also it registers acceptance result. In static check, the system checks file type and file size, and compile the submission file (Fig.5). When learner submits binary file or danger file potentially malicious, the system excludes this file. Danger file includes some functions such as OS command execution, input work from keyboard. The result of static check is notified to students regardless success or failure. In dynamic check, the system performs checking battle as testing for compiled strategy binary. the binary is tested whether it ends correctly by battle with sample strategy. The system excludes some strategies such as stopping by error, causing endless loop. If checking battle finishes successfully, it registers strategy DB as a strategy which can participate preliminary league.

Subsequently, in battle processing, the submission strategy performs battles with all submitted strategies and it calculates weighted winning grade (WWG) (Fig. 6). First, the system gets strategies in the league, and creates tournament chart with them. Next the system performs battle by tournament chart. In this time, battle results and game records are saved as temporary file. After finishing battles, it registers opponent ID and battle result to battle DB. game records are XML file, so the system saves saved file path to battle DB. Then, it calculates temporary WWG by battle results.

4.4 Introduction of time of suspension of submission

Actually, it takes considerable time for battle processing as submissions increase. In addition, WWG is affected by WWG of other strategies and it also affects. So, processing becomes complicated when the submission of strategies is successive. WWG also changes frequently. It sometimes happened system fault or significant delay of processing on legacy system. For the reason, we deal with thinning round-robin series for reducing number of processing.

So, in 2016, we decided to stop strategy submission from 24 to 6. During this period, we finish battle processing every day. The battle result at that time, the system calculates determinately WWG at the day. These results are reflected to ranking, and published to Web at next morning. In new system, battle processing and database management were improved, so system fault or significant delay were not happened. However, there is still room of improvement in the immediacy of the result reflection.

4.5 Reconstitution of database

In legacy system, we use XML as database. However, when system registers each battle results about 1000 strategies, it takes a lot of time to load from database. So, we reconstitute

database using RDBMS to accelerate. We use PostgreSQL 9.6 as DBMS.
Table 1 shows a list of tables composing database. Submission table "Submission" registers all the submitted files. It is management as submission log for learners. strategy table "Strategy" registers only strategies which is pass to acceptance processing and participate in the preliminary league. ID is given to strategy with serial number for each learner. It uses for making tournament chart.
Battle result table "Battle" saves all actual battles. It registers self and opponent ID, battle result, and path of the battle record file. battle record files are still in XML format for use in record replay module on WinG-LA. In the future, we consider migrating to file saving as JSON format for replaying on WinG-CS.

4.6 Use of Docker in battle environment

We describe using Docker on WinG-CS. On WinG-CS, we register Docker image as a template of execution container in advance [14][15]. This image includes openJDK, and be saved execution library and commands. In executions of individual battle, WinG-CS creates execution container based on the image. Execution container includes tournament chart (Fig.7).
And also, execution container mounts strategy saving directories and game record directories on the host machine. As the result, we can directly read and write data on host machine from container. First, the container starts battle processing according to execution commands. After battles, the container transfers all battle records to host machine, and also any databases are updated. Finally, the container is discarded. In such approach, it also facilitates load balancing using multiple machines.

4.7 Improvement of GUI in the student side

We describe GUI of the improved system. In top page of the league, it shows tables of strategy ranking and individual submission history by tab. Learners can be anonym by nickname as player. General ranking table in Fig.8(a) shows ranking of all submitted strategies so far. Rows of index strategies and their own strategies is changed color, and emphasized from other learner's strategies. Submission history tab in Fig.8(b) shows only own strategies. Submitted files are showed in time series. If user clicks strategy name, they move to strategy detail page.
Strategy detail page shows detailed information of individual strategy in various tabs. strategy summary tab in Fig.8(c) shows, conventional information such as submission date, battle result information, and strategy code. Learners select their own the best strategy for final league. Records list tab in Fig.7(d) shows all results of the strategy. Learners can download each battle records, and replay on WinG-LA. They also can download all battle records at once.

Table 1. Database structure

User	Student ID, Login information
League	Subject's name, Contest period
Player	Player name, Role, Best strategy ID
Submission	Strategy name, Check status
Strategy	Submission ID, WWG, File path
Battle	Strategy ID, Battle result, File Path

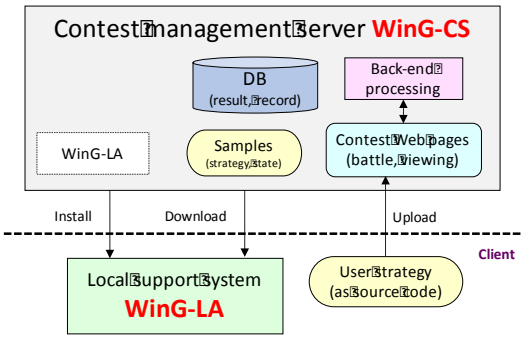


Figure 3. Architectures of support system WinG

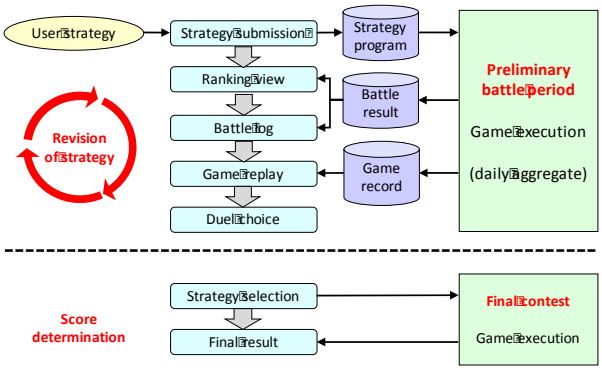


Figure 4. Functions of WinG-CS

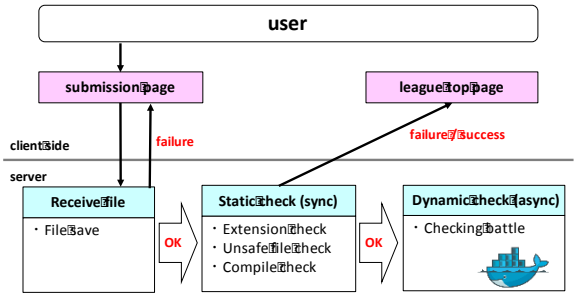


Figure 5. Flow of acceptance processing

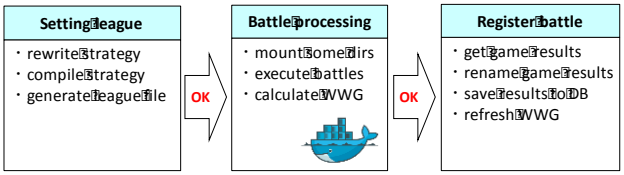


Figure 6. Flow of battle processing

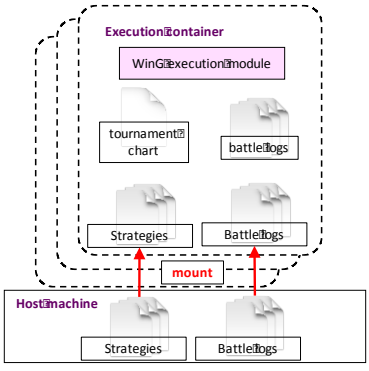
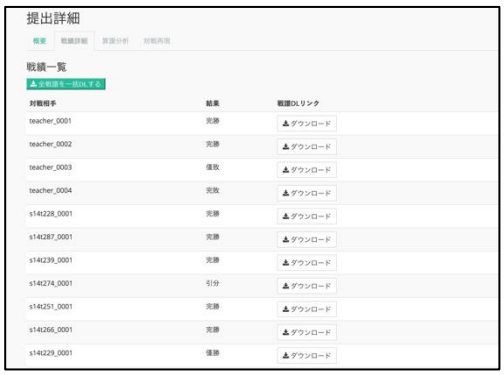


Figure 7. Structure of exrcution environment in Docker container

(a) Total ranking tab

(b) Submission log tab

(c) Strategy summary tab



(d) Record list tab

Figure 8. GUI of student side

ISBN: 978-1-941968-42-0 ©2017 SDIWC

5 IMPROVEMENT OF RATING INDEX "WWG"

5.1 Competition method and winning point

In this exercise, a match consists of two game sets, with players taking it in turns to go first and second for fairness. A player is given winning points by winning or losing. Table 2 shows winning point (WP). If the result of two sets is one win and one loss, the outcome is decided on points based on the sum of captured stones. By WP, we can subdivide results which judged draw simply, and classify clearly. If battle numbers are same in round-robin tournament, individual total sum of WP stands for a strength of a strategy. Simple winning grade SWG, which is calculated dividing WP by 4 times of battle number, is actual value on the interval [0,1]. SWG is a percentage of victory considering WP. If the result is composed only complete win and complete lose, SWG agrees normal percentage of victory.

5.2 Necessity of rating index

However, in the preliminary league, the ranking may be influenced by a tendency of the main strategy group during a period. At the beginning of the period, very weak strategies by incomplete programs are included. A high winning degree by these does not show exact strength. While a few aggressive students upload many similar strategies, deviation of strategies causes inaccurate or unfair result with ill-suited battles. So, the affinity for same type of strategy strongly appears. In some cases, it is possible to prepare dumping strategy like kingmaker, and to raise the ranking of the strategy which wants to win. So, it is necessary to reduce bias due to apparent strength and affinities. For this problem, winning point grade considering strength of opponent is necessary [16].

5.3 Improvement of weighted winning grade WWG

In this exercise, we introduce weighted winning grade ("WWG") as rating index [17]. WWG is refinement index of SWG. It can make comparisons even in situation with different number of battles in the middle of preliminary league. WWG is the actual value on the interval [0, 1], and weights own wining point by opponent's winning point. It is necessary to calculate recursively because own WWG changes with the change of the opponent WWG. When the number of battles is small in early stage of preliminary league, definition of WWG until 2016 has a defect that the fluctuation of WWG is large. For that reason, we adopt a new definition from 2017. In addition, at a sufficient number of battles, the deference between the new and the old is not noticeable except for extreme battle results.

Here. fight number of strategy x is $N(x)$, winning points for strategy x of y is $WP(x, y)$, let $WWG_k(x)$ be the WWG for strategy x by k-th calculation. Initial value is $WWG_0(x) = 1.0$. $WWG_{k+1}(x)$ is calculated by recurrence formula. We repeat the calculation of the formula until the change of the value fails below an appropriate threshold. The value is confirmed WWG.

$$WWG_{k+1}(x) = \frac{0.5 + \sum_{i \neq x} \left\{ WWG_k(i) \times \frac{WP(x,i)}{4} \right\}}{\sum_i \{WWG_k(i)\}}$$

In this definition, it includes virtual battle with myself. +0.5 in the numerator means to treat it as draw. Here, 1st calculation $WWG_1(x)$ is simple winning grade $SWG(x)$ including self-draw. And also, denominator is sum total of WWG, and it makes WWG relative. If a strategy wins to all strategies, WWG doesn't result in 1.0 by self-draw. As the count of complete win increases, approaches slowly to 1.0. And also, WWG when complete loss to all strategies is not 0.0 but a half of the reciprocal of sum total of WWG of all strategies by self-draw. The more number of battles, approaches slowly to 0.0.

Previous definition of WWG doesn't include self-battle to a numerator. In this way, if there is all-loss strategy, weight of win to the strategy is 0.0. Then, it spreads in recurrence formula, and all WWG converge to 0.0. On early stage in 2016, the trend becomes noticeable, so we introduce correction of the normalization. However, on the contrary, if all the forces are

in equilibrium and the victory or defeat becomes close to half, the WWG may be contrary to intuition. By new definition of WWG, it can calculate reasonable value in extreme situation of battle results. And also, convergence is improved.

Table 3 shows winning or losing in 6 strategies of A to F and convergence condition of WWG. (a) is the case when superiors always win completely to subordinates. A with all winning converges to 0.866, and F with all defeat converges to 0.183. Other strategies also converge to a reasonable value in about 6 times with vibrating up and down. On the other hands, (b) is a little more complex winning and losing situation. A and B are 4 wins and 1 loss. However, A wins B addition to D, E, and F. B wins C addition to them. At that time, A wins stronger opponent than B wins, so A is higher WWG than B. And also, E and F is 1 win and 4 losses. However, E wins only D, and F wins only E. E wins higher rank strategy than F. So, E's WWG is higher than F. Both C in (a) and C in (b) are 3wins and 2 loses, but WWG of C in (b) is a little higher than C in (b) because of win to the top.

Table 2. Winning point

2 wins	complete win	+4
1 win 1 lose	decision win on point (more captured)	+3
1 win 1 lose	draw by the same point (same captured)	+2
1 win 1 lose	decision lose on point (less captured)	+1
2 loses	complete lose	0

Table 3. Standing and calculation status of WWG

(a)			Win	1	2	3	4	5	6	7
A	5	0	BCDEF	0.917	0.861	0.857	0.869	0.867	0.865	0.866
B	4	1	CDEF	0.750	0.611	0.615	0.643	0.636	0.633	0.635
C	3	2	DEF	0.583	0.417	0.451	0.478	0.464	0.462	0.465
D	2	3	EF	0.417	0.278	0.341	0.353	0.338	0.339	0.341
E	1	4	F	0.250	0.194	0.264	0.256	0.246	0.249	0.250
F	0	5		0.083	0.167	0.198	0.183	0.180	0.183	0.183

(b)			Win	1	2	3	4	5	6	7
A	4	1	BDEF	0.750	0.722	0.729	0.739	0.740	0.736	0.737
B	4	1	CDEF	0.750	0.667	0.701	0.706	0.698	0.699	0.701
C	3	2	AEF	0.583	0.583	0.598	0.574	0.584	0.587	0.583
D	2	3	CF	0.417	0.444	0.549	0.520	0.505	0.516	0.516
E	1	4	D	0.250	0.306	0.318	0.331	0.326	0.322	0.325
F	1	4	E	0.250	0.250	0.271	0.258	0.266	0.265	0.263

6 OUTLINE AND RESULT OF PRACTICES

6.1 The outline of practices in our programming exercises

Table 4 shows summary of exercise practices from 2013 to 2016. Exercise period for this practice is about substantially 5 weeks. In first lesson, we introduce the rule of game, and also explain strategy programming, local execution environment, and contest management server. Around next week, as preliminary contest period they approach in extra-curriculum. In a class, we check general ranking or submission state between other practice, and explain additionally. The deadline is the last of preliminary league. After the time, learners select the best strategy for final league. In 2014, server had failed. In 2015, end time of other practice was postponed, so start time of this practice was delayed. For this reason, we extended the practice period because of conflicting the deadline of other practice. However, the substantial development period is as usual. In 2016, the start of contest was also delayed in relation to other practice. We presented development environment before in 2 weeks before the start of contest. It is as short as preliminary league. So, we also substitutional development period as usual. In an exercise in 2016, we had not adopted thinning round-robin series, so submitted strategy battles with all registered strategies at that time. It is no problem until the middle of contest period by improvement of battle processing. But, in the end stage of the league, delay was happened by increasing number of submissions.

6.2 The transition situation of submission

In 2016, Total submission number is 800. The average of submissions per person is about 20. The maximum submission number is 67. It greatly exceeded in 2015. Fig.9 shows a transition of submission number in each year. Transition in 2016 similar to it in 2013. Plateau in the middle of contest period is less than other year, number of submissions constantly increased. Fig.10 shows a frequency distribution of submission. In 2013, There is a

learner who submitted strategies over 100. In 2016, learners who submitted five or less was decreased. And, the distribution peak is on the right compared to normal year. On the other hand, there are a half of the learners who submitted 20 or less. In the future, it is necessary for these learners to support.
Currently, we focus on transition of result by correlation of "STG" and WWG. It is necessary to find low-rank students considering submissions, or less submissions and low-rank students, and support or encourage them. To defect these cases, we use score transition graph STG [17]. On the other hand, we also induce high-ranking students considering less submissions to work continuously. If it is difficult to develop new strategy, we examine to induce learners to change their policy to quality of code.

6.3 Application of improved WWG to final leagues

We tentatively apply the new definition of WWG in chapter 5 to final leagues after 2013. Fig.11 shows their correlation. The horizontal axis X is SWG, the vertical axis Y is WWG. Spearman's rank-order correlation ρ is also mentioned in these figures.
Each year, correlation ρ exceeds 0.99, and there is strong correlation between WWG and SWG. So, it is considered to calculate ranking not against intuition. Also, strategies with 0.4 to 0.7 SWG are differentiated by strength of opponents. on the other hand, lower difference is shrinking. This is because raising by influence of self-battle.

6.4 Improvement of thinning round-robin series

From 2017, we improve a method of thinning round-robin series, and we realize to publish result efficiently and instantaneously thinning round-robin series. First, after strategy submission, a submitted file performs checking battle with sample strategy. If there is no such as run-time error, it is registered as a strategy to participate in preliminary league. In preliminary league, the strategy performs provision battle at first. In provision battle, it meets about 10 prepared index strategies, in battle game. WWG of Index strategies are calculated as criterion in advance. After finishing provision battle, the system calculates temporary WWG of the strategy, and publishes general ranking table as a flush report.
Next, during the time of suspension of submission from 24 to 6, the system performs entry battle for all submitted strategies while the day. In entry battle, a strategy meets about 100 strategies which submitted by the day before in a battle. At this time, the system equally chooses strategies from WWG distribution. By this result, the system recalculates all strategies' WWG, and publishes general ranking table as daily result in early in the morning of the next day. This is initial WWG for strategies submitted the day before. After next day, strategies perform defensive battle as an opponent of other strategies in provision battle. Strategies submitted in early stages of the contest are performed more defensive battle. Ordinary WWG of strategies change by daily defensive battle.

7 CONCLUSIONS

We have practiced an applied Java programming exercise with board game strategy. The exercise is contest style. Strategies are battled on server, and published their own result or ranking. He modifies his code repeatedly by the feedback during a contest period. we revise contest management server to improve reliability and efficiency of battle processing. Especially, we adopted virtual environment, and make battle process efficient. We also revise rating index WWG to refine battle results. And also, we reconsider methods of thinning round-robin series by the result in 2016. Thus, we arrange 3 stages of battle situation. In addition, we laid the foundations for distributed processing and thinning round-robin series.
In future work, we try to improve reliability and efficiency. First, we implement functions about revised thinning round-robin series to improve efficiency. Next, we introduce code metrics as an internal evaluation to make students aware of not only battle result but also quality of code. It has learners promote continuous

modification for refactoring. In addition, we try to analyze learners' action during contest.

ACKNOWLEDGEMENT

The authors would like to thank students who participated in this exercise, and Dr. Takahashi and Mr. Okazaki who assisted this exercise.

Table 4. summary of practices in each year

year	2013	2014	2015	2016
Period (week)	7	7	6	5
Learners	37	35	44	40
WWG	○	○	○	◎
Thinning round-robbin series	◎	○	○	×
Total of submission	942	377	406	800
Average of personal submission	25.5	10.8	9.2	20.0
Maximum of personal submission	142	32	44	67

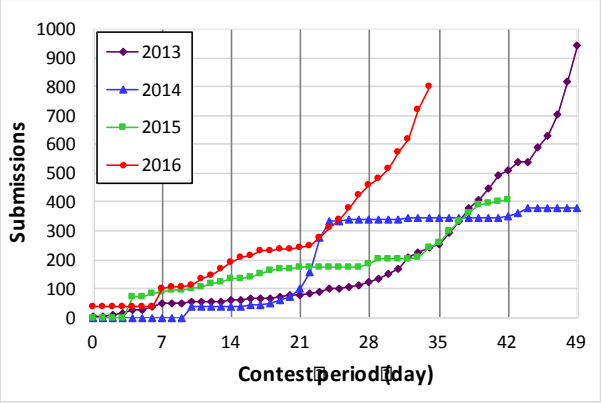


Figure 9. Ogive curves of submission number in each year

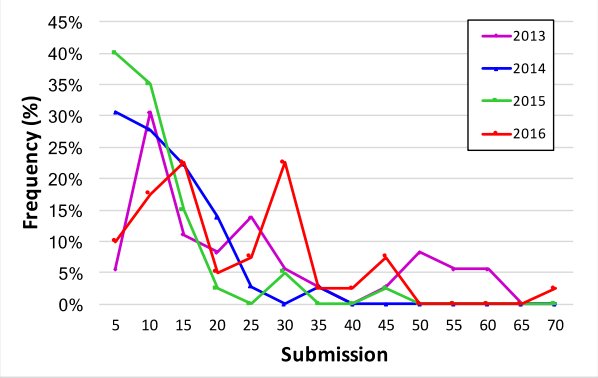


Figure 10. Frequency distribution of submissions

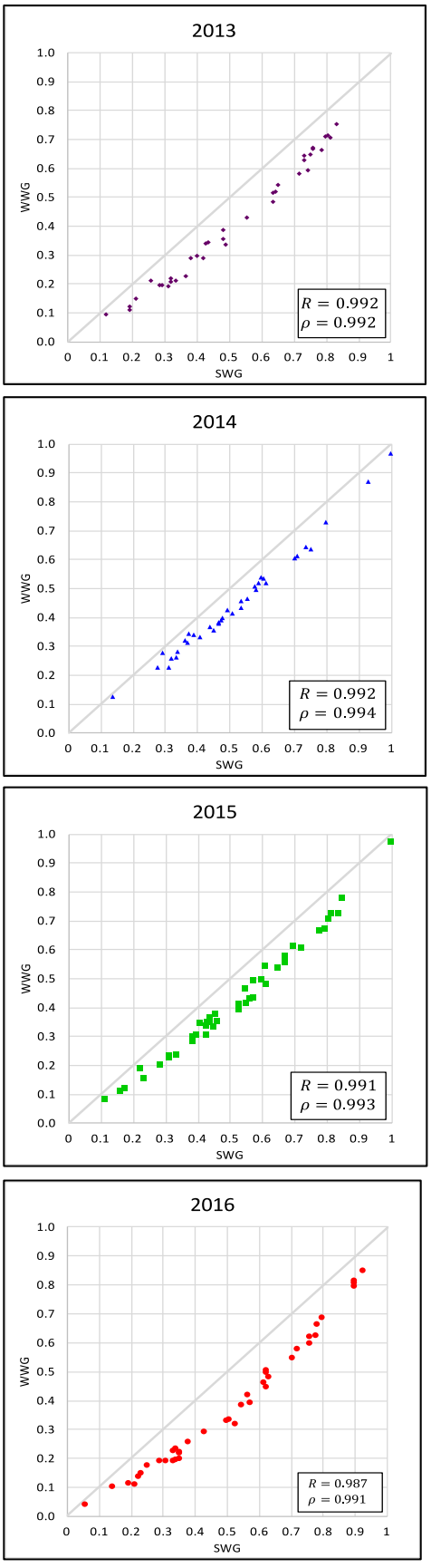


Figure 11. Application of improved WWG to final leagues in each year

REFERENCES

[1] H. Ozaki, H. Tominaga, T. Hayashi, and T. Yamasaki, "Support System for Java Exercise with Strategy Programming about Board-Game Gogo", Proceedings of ITHET 2007, pp.530-535, 2007.

[2] K. Yamada, and H. Tominaga, "Support System WinG and Applied Programming Exercise with Board-Game Strategy", Proceedings of ITHET 2012, No.PS9, pp.1-6, June 2012.

[3] Pente Net, Pente Net, http://www.pente.net/

[4] K. Yamada, and H. Tominaga, "Ranking Analysis of Battle Result of Board Game Strategy in Java Programming Exercise", Proceedings of WCTP 2013, No.6, pp.173-184, September 2013.

[5] F. Gemba, N. Hanakawa, and H. Tominaga, "Educational Approach and Practices for an Applied C Programming Exercise with Poker Card Game Strategy and a Contest Style", Proceedings of ICIA 2016, Vol.2016, pp.97-104, November 2016.

[6] IPSJ, SamrAI Coding, http://samuraicoding.info/

[7] IEEE CIG, Computational Intelligence & Games, http://www.cig2017.com/competitions-cig-2017/

[8] MIT, BATTLECODE, https://www.battlecode.org

[9] CodinGame, CodinGame, https://www.codingame.com/start

[10] J.C. Canada, T.J. M. Sanguino, J.J.M. Guervos, and V.M.R. Santos, "Open classroom: enhancing student achievement on artificial intelligence through an international online competition", Journal of Computer Assisted Learning, Vol.31, No.1, pp.14-31, 2015.

[11] D.M. Yoon, and K.J. Kim, "Challenges and Opportunities in Game Artificial Intelligence Education Using Angry Birds", IEEE Access, Vol.3, pp.793-804, 2015.

[12] A. Bezgodov, A. Karsakov, K. Mukhina, D. Egorov, and A. Zakharchuk, "Learning AI techniques through bot programming for a turn-based strategy game", Proceedings of European Conference on Games-based Learning, Vol.2015-January, pp.66-74, 2015.

[13] J. DeNero, and D. Klein, "Teaching introductory artificial intelligence with Pac-Man", Proceedings of the Symposium on Educational Advances in Artificial Intelligence, pp.1-5, 2010.

[14] Docker Inc., Docker, https://www.docker.com/

[15] K. Jiang, and Q. Song, "A Preliminary Investigation of Container-Based Virtualization in Information Technology Education", Proceedings of the 16th Annual Conference on Information Technology Education, pp. 149-152, September 2015.

[16] A.N. Langville, and C.D. Meyer, Who's# 1?: the science of rating and ranking. Princeton University Press, 2012.

[17] N. Hanakawa, F. Gemba, and H. Tominaga, "Score Transition of Card Game Strategy as Personal Progress Situation in an Applied C Programming Exercise with a Contest Style", Proceedings of WCTP 2016, Vol.6, pp.140-151, September 2016.

On Trust Management Framework in Video Streaming Applications Over Mobile Ad Hoc Networks

Thulani Phakathi, Francis Lugayizi, B. Esiefarhenrhe, Bassey Isong
Department of Computer Science, North-West University
Private Bag x2046, Mmabatho, Mafikeng, South Africa
katshego@gmail.com{francis.lugayizi, bassey.isong, 25840525(@nwu.ac.za)}

***Abstract*— Trust is an important computing networks concept and remains an issue not just in social platforms but also in networks in the deliverance of Quality of service (QoS). Video Streaming has, over the years gained prominence in mobile and vehicular ad hoc networks. However, it is faced with several challenges of being vulnerable to different attacks and poor QoS. Therefore, this paper proposed a robust and efficient trust management framework in an effort to eliminate the poor QoS due to network detriments caused by the dynamic topology of a Mobile Ad Hoc Network (MANET). The proposed trust framework in the network is aimed to assist in node accountability and QoS attainment. That is, it does not remove the dynamic topology issue but it strives to attain nodes' trustworthiness and excellent QoS despite the conditions set out against the network. The proposed trust framework is the application-centric trust management framework with distributed trust computations (AppTrusFram). It merges the concept of trust together with QoS in an application scenario. The paper present a theoretical-design solution to the topology related issues and discussed issues found in other proposed solutions.**

Keywords*— *Routing protocols, MANETs, Trust framework, Video streaming, QoS.

1. INTRODUCTION

Mobile Ad Hoc Networks (MANETs) also referred to as an infrastructure-less network [4] is a network technology that has gained significant attention in the research world in recent years due to related protocols challenges it faced. MANETs are an emerging technology which offer network users interactions without any central infrastructure irrespective of the geographical location of the users. MANETs have been an active area of research for the last few years and their growth is promoted by the growing need to provide the users with the network support at their own convenience [5]. The network is primarily useful in military and other tactical applications such as emergency rescue or exploration missions [6]. Their commercial success has been due to advances in wireless technology and several standards have been developed for routing in MANETs. The Internet Engineering Task Force (IETF) are responsible for regulating the new group for MANETs and IEEE standard 802.11 has contributed to research interest done with MANETs [4].What also facilitated the explosive growth is the continued production of smaller and faster devices which makes MANETs the fastest growing network.

Today, MANETs are considered as a household technology service in the mobile network industry. Its growth has been commended globally and has attracted so much attention in recent years with the invention of Vehicular Ad Hoc Networks (VANETs) and its integration into the automotive industry. In particular, the fast demand for video streaming applications like video conferencing and video-on-demand are central to MANET technology. Researchers in recent studies have shown great interest in quality of service (QoS) in the mobile network realm.

QoS is considered a set of service requirements that a network is required to meet in the movement of packet streams from the source to destination. With the explosive demand of QoS provisioning for evolving applications (e.g. video and voice), it is proportionally appropriate to ensure that these services are supported in ad hoc networking environments. In the field of telecommunications, QoS was defined as set of requirements on all the aspects of a connection, such as crosstalk, response time, loss, echo, signal-to-noise ratio, interrupts, frequency response, loudness levels, and so on. Moreover, QoS constitute the ability to proffer a different level of priority to different users, applications, or data flows, or to ascertain a precise performance level to a data flow [8]. However, in

order to achieve QoS, the concept of routing is indispensable. Routing involves information movement in a network from a source to a destination [8]. During routing, at least one node within the network which act as an intermediary is met during the information movement. Thus, routing can be used to determine optimal routing paths as well as the transfer of packets via an inter-network [9]. Moreover, routing may be either dynamic or static [9]. QoS is measured using specific metrics within the network such as bandwidth, jitter, delay, packet loss etc.

In the realm of MANETs, the mobility of nodes, however, is rapid and unpredictable over time. MANETs, like all other wireless networks are more vulnerable to attacks and other weaknesses as compared to wired networks. The limitations in MANETs become especially exacerbated in the multi-hop networks where multimedia streams suffers aggregate effect such as packet drop, propagation delay and jitter of each connected link along an end-to-end path. Due to node mobility, there are frequent route breaks which results in routing updating that is time consuming. Consequently, packets might be lost in bursts for shorter periods of time since they are sent on non-working routes [7] which in turn impacts the QoS adversely.

Therefore, this paper proposes a robust and efficient trust management framework in an effort to eliminate the poor QoS due to network detriments caused by the dynamic topology of MANETs. The proposed trust framework is an application centric-framework which amalgamates the concept of trust together with QoS evaluation. The objective is not to remove the dynamic topology issues but it strives to attain nodes' trustworthiness and excellent QoS despite the conditions set out against the network. The proposed trust framework is called application-centric trust management framework with distributed trust computations (AppTrusFram). Furthermore, we present solutions to the topology related issues provided by the proposed trust model and discussed issues in other proposed solutions.

This paper is organized as follows: Section II discusses the various routing protocols (RPs) in MANETs, Section III is on trust in MANETs and Section IV presents trust management frameworks studies while Section V presents the proposed framework.

2. ROUTING PROTOCOLS IN MANETS

This study focuses on reactive and proactive protocols which are the RPs that are based on topological information in MANETs as well as in VANETs [3][19][22]. Each is discussed as follows.

2.1 Proactive Protocols

These protocols operates by periodically trading control messages of known routes between all the network routers [5][6] [35]. Proactive routing protocols include DSDV, OLSR, Fishy State Routing (FSR) and so on. Thus, OLSR is the focus of this study.

a)OLSR protocol: In the OLSR protocol, every station found in the network choses a neighboring nodes set known as the multi-point relays (MPR), which rebroadcasts the packets in turn. To this end, neighboring nodes identified not found in the MPR set has the capability to only read and process the packets [5]. Moreover, OLSR retains tracks in path finding table in order to proffer a route if necessary [8].

2.2 Reactive Protocols

These RP are also called on-demand protocols and include AODV, DSR and TORA [22].

a)DSR protocol: In DSR, the discovery of route begins on-demand and the whole path to destination are placed on the routing table other than using next hop node as in the AODV. The packet header has the address of all the nodes in transition needed by the packet to get to the destination node [10]. In addition, nodes can be dynamically discovered in a source route through complex networks hop to any terminus in VANETs by DSR. However, the protocol lacks the mechanism to identify unstable routes leading to forwarding to data packet to broken link [11]. In operations, DSR implements three unique techniques for controlling packets for path discovery and maintenance [10].

b) AODV protocol: AODV operates on a pattern known as hop-by-hop and uses flat routing tables having one entry per destination [11]. Unlike DSR, the AODV algorithm enables high mobility, dynamic, multi-hup and self-initiating routing

among collaborating vehicles that are interested in ad-hoc network establishment and maintenance. In this case, routes can be acquired speedily for new destinations by mobile nodes without keeping routes which commination is not active to destination.

c)TORA protocol: TORA protocol works on controlling the propagation of messages in the ever changing ad-hoc networks. Node are known to clearly begin a query only when data is to be sent to a destination. In terms of performances, TORA performs much better as compared to that of DSR in a network [4]. According to Qureshi and Abdullah [4], the reason be due to the protocol's ability to minimize the communication overhead in a dynamic environment thus, making it more reliable for changing Ad-hoc networks. In addition, Palta and Goyal [9] stated that one good aspect in the design of TORA is that there is localization of control messages to few node sets and the effective way link failures are handled.

3. TRUST IN MANETS

In this section, we present the nature of trust and related works on trust management in MANETs.

3.1 Trust Factor

The term "trust" is used in the perspective of how one party (normally referred to as trustor) is willing to rely on another party (may be referred to as trustee). It is greatly attributed to human beings and their relationships in social groups. However, in the Computer Science discipline, a trusted component has elements within itself that another component within the system can rely on. For instance, if component **A** trusts component **B**, this means that any violation of properties found in component **B**, will ultimately affect the correct operation of **A** - *dependency*. This, however, doesn't mean if **A** trusts **B**, then component **C** can trust **B**. In the realm of MANETs, due to its characteristics, trust management is required for participating nodes that communicate together in a network to provide a satisfactory or acceptable level of trust relationships among them without any previous interactions [1]. Nevertheless, trust management in MANETs are enormously challenging because of its dynamic nature and characteristics which is attributed to topology changes as well as uncertainty and incompleteness.

a)Direct Trust: In Kiefhaber *et al.* [20], direct trust is a form of trust which involve the experiences an entity has created directly with another entity it interacts with which is computed using trust value. Typically, trust values are computed using the either the mean or weighted mean of previous experiences. Direct trust is application-centric or rather specific. The application has the decision to determine whether an interaction made by one entity to another was successful. Direct trust was chosen based on the merits that it is reliable in terms of rankings from confidence trust and reputation[20]. On the middleware level, which in this case is the application server, a node's reliability can be computed through observation of the message flow in the distributed system. initially, the Delayed-Ack mechanism was used and was later changed.

b)Topology Constrainsts in MANETs : A network topology that is trustwory must be assured via the use of robust routing protocols in the ad hoc networking stream [1]. They are required because of the frequent routing updates caused by the dynamic nature and characteristics existing in MANETs. Providing QoS at a better scale in such environments is a huge challenge [4]. The existing stochastic nature of MANETs' quality of communications poses challenges in offering concrete guarantees on the applications computed in mobile devices. Thus, a QoS that is adaptive must be realized coupled with a strong trust relationship framework over the traditional methods reserving resesources to support multimedia streaming services. But, due to the constant change in network topology, the issue of routing packets between nodes poses a great problem. Multicast routing also poses a challenge since the multicast tree ceases to be static since nodes in the network move randomly. The routes between nodes often have multiple hops which is considered complex than its single counterpart. In MANETs, the nodes are mobile and this feature could result in nodes getting out of range within the network [4]. This causes frequent loss of links between nodes. That is, when nodes are in motion, the state of present node links are most susceptible to changes, or possibly break. Re-routing is an alternative routes that are broken.

c)Trust and QoS: Trust and QoS have almost similar metrics in order to evaluate their efficiency. There are different ways to define trust. Trust is considered as a directional relationship between two entities which is critical relationship building a relationship between them in a network [17]. Though trust has been considered as a computational model, it is viewed differetly by different research communities. QoS in MANET [8] which is universally growing area. With the rapid advanvement in multimedia technology today, there is urgent need for mobile technology and real time applications to strictly support QoS such as throughput, delay, energy consumption, jitter etc. Trust and QoS share similar metrics e.g. delay, throughput and packet dropping rate. The correlation between these concepts exists not only through shared performance metrics but also the assurance of good service in the network for the end user. QoS and Trust is what stands in between the network and the Applications/Users. It is not easy to provide QoS support without having the right QoS requirements. A particular level of trust must be established and that is, only trustworthy nodes that perform as desired will participate. Trust is dynamic [16] just like MANETs which have an ever changing topology state. This means that the trust value should be based on temporary and yet local information. In addition to this, the trust value is different for similar nodes is different. This is influenced by that each node goes through different situations in terms of the dynamic topology.

4. RELATED WORKS ON TRUST MANAGEMENT FRAMEWORKS IN MANETS

Trust remains [1] a relevant subject within the research field and continues to attract interest from network analysts and developers. The notion of using trust to eradicate security problems in networks has also been relevant in the research field. Ferdous *et al.* [12], proposed a novel approach to address problems by using trust as a metric. Their approach is based on the node responsibility to build an acceptable trust level and monitoring [12]. Their work is based only at node level and this paper seeks to go deeper into considering other factors like the QoS at the network and trust in the entire network. This means that a bridge between trust and QoS will be built in this paper.

Sen [17] proposed a reputation and trust-based security framework for MANETs that detects packet-dropping attacks launched by malicious and selfish nodes. The framework was based on a trust model which is based on the reputation of other network nodes. The study stated that MANETs are prone to security threats in which a node could hid its initial identity and disguisedly re-enter a network environments where users are penalized for behavior that seems selfish or malicious. The solutions proposed involved invoking a univocal relation between a terminal and the initial identity it assumed when it first enters the network [17]. The work was implemented for only small scale performances and not high scale. It was at a simulation area of 100*100m. DSR protocol is widely known for its scalability problem especially when the ad hoc network topology varies. Different results could have been achieved through the OLSR protocol or rather the TORA protocol which can evaluate either a proactive or even a reactive protocol.

Li *et al.* [14] also proposed a trust-based framework which can be incorporated with diverse single-copy data forwarding protocols in Opportunistic Networks (OppNets). It aimed at carrying out an in-depth assessment of the potential encountered for data delivery [18]. Their work aimed at counteracting arbitrarily forwarding attack [18]. Zhang *et al.* [19] also focused on the problem of control delay-constrained topology having in mind, other problems like account delay and interference. The study proposed a cross-layer distributed algorithm known as interference-based topology control algorithm for delay-constrained (ITCD) MANETs [18] while taking the interference and delay constraints into consideration. Additionally, the study investigated the effect of node mobility on the interference-based topology control algorithm where any node considered not stable is removed. The results obtained from the simulation performed showed that ITCD reduce the delay and in turn also improve the performance effectively in delay-constrained mobile ad hoc networks.

Li and Kato [11] proposed an Objective Trust Management Framework (OTMF).The framework assessed nodes' trust and was used to compel nodes

to collaborate in a manner that is normal. The framework was geared towards designing a robust and attack-resistant trust management framework to overcome vulnerabilities problems in the future. This vulnerabilities include not only topology-related or scalability-related vulnerabilities. Moreover, Shabut et al. [20] proposed a trust model that is recommendation-based. It has a defence mechanism that can dynamically filter out attacks using clustering techniques and the model was empirically evaluated. The empirical analysis demonstrated the attributes of robustness and accuracy in a MANET dynamic environment. However, the results, cannot be validated in an experimental process since the framework is based on recommendation. Thus, our proposed framework is one based on direct trust

Trust is an important feature in networks [10]. The nature of MANETs still make the guarantee of efficient trust a complex task due its highly dynamic topology constraints. The emergence of MANETs calls for the addressing of many problems perceived in networks and to also optimizing some of the existing network resources [2]. The question that remains unresolved is "c*an trust be truly guaranteed in a MANETs*?" The answer to this question according to this study will be based on the QoS. Different authors have presented different trust frameworks. According to Li and Kato [11], existing trust frameworks are faced with great challenges under hostile environments, which can adversely affects their performance. This means that the reason most frameworks failed is that they did not address the problem of topology and its being dynamic in these type of networks hence the need for robust frameworks that will be resistant and still get to give out optimum performance.

Ferdous *et al.* [12], also developed a simple node-based trust management technique for MANETs. It provides multiple standpoints of trust, its properties which are trust metrics, and the insights into the customization the metrics meet the requirements and goals of the network trust management (NTM) scheme. Their future work was set to develop trust management mechanisms having the required attributes like scalability, adaptation to topology dynamics, and reliability [12]. A good scheme is one that encompasses all these characteristics in order to ensure trust in the network. Dynamically changing topology is one of the characteristics and limitations of MANETs. The most important aspect is achieving good QoS.

Thus, a robust and efficient trust management framework is needed to ensure nodes are trustworthy and prevent them from being selfish or performing selfishly as well as provide good QoS. This work seeks to close that gap by building a framework that not only looks at trust but also with reference to video streaming applications in MANETs. Much concentration will be based on the achievement of trust between nodes of a network. Different network scenarios will be established, which means an outlook on performances, behavior and together with their quality of service of video streamed applications will be closely analyzed.

5. PROPOSED APPLICATION-CENTRIC TRUST FRAMEWORK

This section present a high-level overview of the proposed trust management framework and the different components or technologies associated with it. It also present assumptions that we made in connection with the framework.

5.1 System Overview

In this paper, we proposed an application-centric trust management framework with distributed trust computations (AppTrusFram). AppTrusFam is a video streaming application, a rising technology innovation in MANETs which may come in the form of low resolution video or high resolution video. The proposed framework involves application server together with distributed trust computations. (See Fig. 1) The benefits is that it is a management system that is able to compute the end result which is QoS evaluation in the form of throughput, delay, and jitter and so on. In Fig. 1, the AppTrusFam involves neighboring nodes direct trust. The application server plays an important role in the attainment of excellent QoS and is built-in in every workstation since they acts as its own router. This is because it encompasses features that involves security, diagnostics and clustering. Distributed trust computations are based on 'knowledge', 'recommendation' and 'experience'.

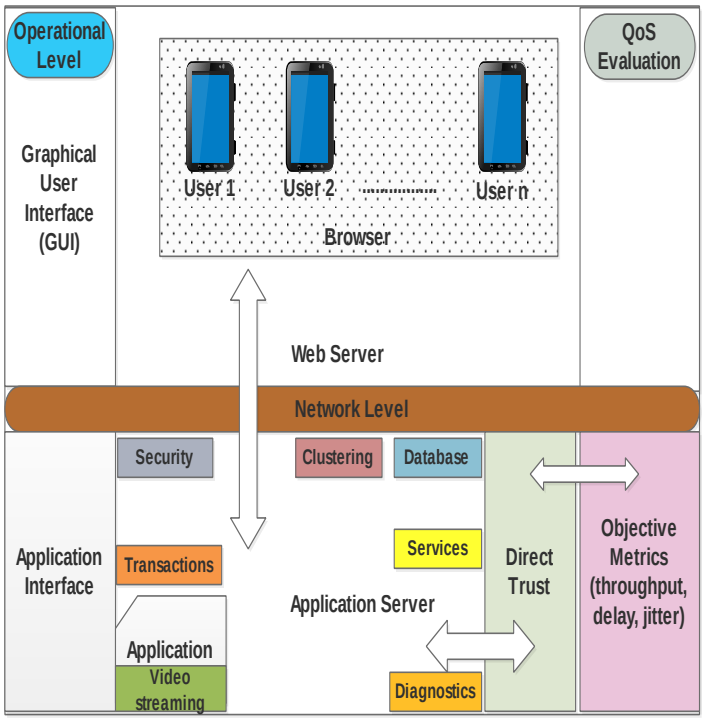


Fig. 1. Trust management architecure

The choice of direct trust is due to its versatility and its ability to sense neighboring nodes since MANET stations operate through WiFi or Bluetooth connectivity. On the other hand, the web-server is responsible for processing client requests and send responses via the http protocol. Within the web-server lies the web browser, an application, which is responsible for retrieving, traversing and presenting http requests onto stream(able) media in the World Wide Web.

6. APPTRUSFRAM COMPONENTS

AppTrusFram shown in Fig. 1 is video streaming which shows a link between direct trust and QoS metrics. The components are discussed as follows:

a)Application server: The Application server consist of the business logic and the application programs using various protocols. It also take charges of all application operations that exist between users and an organisation. It is able to deploy applications and it primarily acts as a middleware, an abstraction and serves that facilitate the design, development and integration of distributed applications in heterogeneous net-environments. In MANETs, the nodes are self-configuring, basically meaning they have their own internal application server.

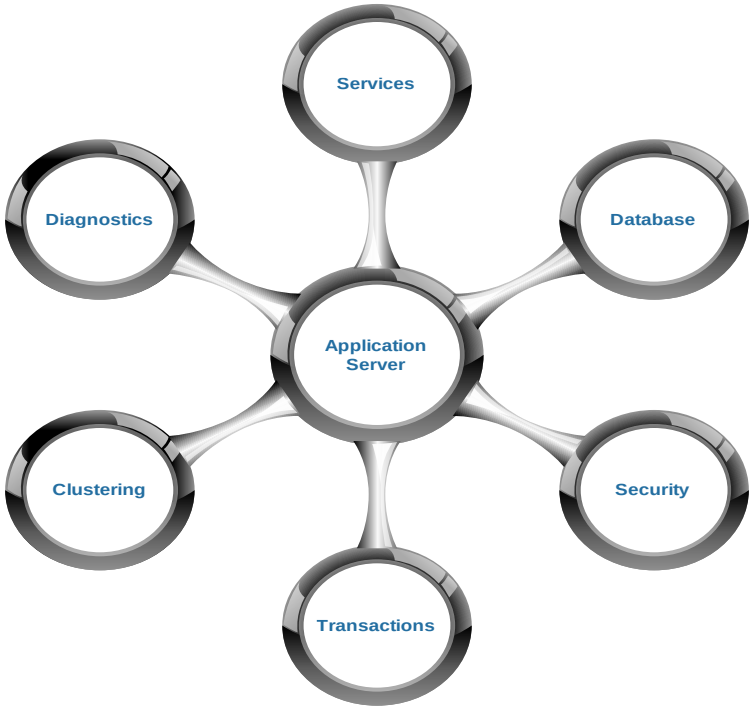


Fig. 2. Application Server functions

Fig. 2 shows the functions of the application server in a breakdown. The application server has a diagnostics attribute to track down and resolve errors. Security is a critical aspect in terms of trust management. This is to make sure malicious nodes to not end derailing performance of the entire network. In the context of this paper, trust among nodes is compromised when one node withholds packets for a long time since transmission is through hops from one node to the other. Another active functionality of the application server is clustering. Clustering ensures fault tolerance of the application server. That is, in the event of hardware failure or part of application failure, services will remain running for users. The application server has

a set of common services and also a database for record keeping of all the transactions or rather.

b)Internet network provider: It is important to note the importance of the Internet Service provider (ISP) when evaluatingquality of experience(QoE). Operations taking place at this level of abstraction is not visible to the end-user. The ISP plays a major role because in between application and web level, there has to be internet connectivity.

c)Web server: The web server plays an important role towards the realization of http web pages and connectivity. This technology is mostly on the user interface side. One can argue that it may be used in the evaluation of QoE.The browser is the main driving force behind the realization of the application stream. Basically, the browser translates these http pages into media content e.g video, voice and ftp. Some webservers are found in the application server but this is not common in most application servers.

c)Operational chain: This shows the level of abstraction at which the framework is under. The user interface is where the end-user can manipulate the system and view the entire flow. The network level however is hidden from the end-user. The programmer or rather network specialist can, however, evaluate the performance of the system at this level.Our framework is evaluated from the application interface since we are not mainly interested in the experiences of the user but rather of the application itself in terms of performance.

d)QoS evaluation: QoS is the end-attainment goal of this trust framework and the framework encapsulates such an option. The evaluation of QoS or rather the provision of high QoS is what rates our framework. Different network detriments are observed while the system executes due to topology variations. QoS metrics like throughput, jitter, and delay and received routing packets. The end user cannot determine the QoS of the system because it is at a different abstraction.The QoS results are a true reflection of the trust that exists among entities in the network.If the quality of service is poor,that would mean that the trust as well among nodes is poor.

e) *Direct trust*: Direct trust is application-centric which decides whether an interaction made between one entity and another was successful. Direct trust was chosen based on the merits that it is reliable in terms of rankings from confidence trust and reputation [20]. The framework operates from the point of view that the involved nodes do not necessarily need to have direct experience with all nodes in the network for them to be able to compute a particular trust level about them. In contrast, the trust is based on second hand evidence that is provided by intermediate nodes. In this way, they benefit from the experiences of other nodes in the network. The framework designed is based on trust from interaction experiences. The attainment of good QoS in the network is evidence that the network itself is trustworthy. Good QoS in a dynamic topology network means or rather validates the framework. The direct trust can be computed using the formula:

$$\textit{Direct trust} = (g + (x)/2) / (q + x) \quad \text{----------(1)}$$

where **g** is the time success to say an event happened, **x** is a positive real number, and **q** represents the event transactions. If the first event of the transaction is a success, the direct trust value increases but inverse to this it decreases. In the trust framework, the success rate can also be referred to as the trust value and can be any real number between 0 and +1. In terms of QoS, the metric called delay plays a major role in finding the true trustworthiness of the network. This would mean that a specific node delivers packets on time and thus increased network efficiency.The trust being evaluated is from node to node

7. PROPOSED TRUST FRAMEWORK OPERATIONS

MANET communication play a major role in the attainment of a proper flow of packets within the framework. Firstly, video packets will always move from the sender to the receiving end. The framework vividly shows two different interfaces that are involved, mainly graphical user interface and application interface. The application interface can be abstracted as where the source of the video lies. It should also be noted that so much of bi-communication is involved around the different component parts. Before an application can be sent, the application server as the middleware in this context has to ensure that direct trust is established.

This means that the node has to be inspected in terms of its trustworthiness to service delivery. The application server is responsible for many functions including a database, clustering and as shown in Fig. 2.When the video file is sent through at network level through the assistance of the Internet service provider, it is meant to reach the web server. At this point in time, the video file is at the user interface side. The web server receives it as an http file or page and translates it back into media content (video file).The web server houses the web browsers just like application servers houses applications. Users get to stream the media content. The connection amongst these components remains bi-directional since trust must be maintained. QoS evaluation metrics like throughput, total packets dropped and delay are implemented. Delay is an important QoS feature since it shows if there were selfish nodes or not. Evaluation can be done from the application interface (for QoS).The application interface and the user interface are both utilized the evaluation for QoE.

8. CONCLUSIONS

In this paper, we have presented and discussed trust management in the perspective of QoS in MANETs and proposed a trust framework called application-centric trust management framework with distributed trust computations (AppTrusFram). The essence is to ensure that nodes are trusted throughout their communication to ensure QoS delivery. We presented the architecture of the trust framework and provided explanation of its components and operations. Based on its intended operations, we believed that if adopted for use in the realm of MANETs, it could go a long way to enhance the QoS delivery of video streaming. As a future work, the utmost intention is to implement and evaluate the proposed framework as well as exploring other different trust establishment methods other than direct trust such as recommendation-based or even hybrid methods and so on.

9. ACKNOWLEDGMENTS

This work was supported by the FRC at the NWU-Mafikeng. We express our sincere gratitude and thanks to them as well as our colleagues in the Computer Science department.

10. REFERENCES

[1] S. Semplay, R. Sobti, and V. Mangat, "Review: Trust management in MANETs," *International Journal of Applied Engineering Research*, vol. 7, p. 2012.

[2] D. S. Aarti, "Tyagi,“Study Of Manet: Characteristics, challenges, application and security attacks”," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, pp. 252-257, 2013.

[3] A. Rajaram and D. S. Palaniswami, "A trust based cross layer security protocol for mobile Ad hoc networks," *arXiv preprint arXiv:0911.0503*, 2009.

[4] P. Goyal, V. Parmar, and R. Rishi, "Manet: vulnerabilities, challenges, attacks, application," *IJCEM International Journal of Computational Engineering & Management*, vol. 11, pp. 32-37, 2011.

[5] M. S. I. M. A. Riaz and M. Tariqu, "Performance analysis of the Routing protocols for video Streaming over mobile ad hoc Networks," *International Journal of Computer Networks & Communications*, vol. 4, pp. 133-150, 2012.

[6] G. D. Delgado, V. C. Frías, and M. A. Igartua, "Video-streaming transmission with qos over cross-layered ad hoc networks," in *2006 International Conference on Software in Telecommunications and Computer Networks*, 2006, pp. 102-106.

[7] M. Lindeberg, S. Kristiansen, T. Plagemann, and V. Goebel, "Challenges and techniques for video streaming over mobile ad hoc networks," *Multimedia Systems*, vol. 17, pp. 51-82, 2011.

[8] Y. Singh and M. V. Siwach, "Quality of Service in MANET," *Int. J. Innov. Eng. Technol*, 2012.

[9] N. Sharma, S. Rana, and R. Sharma, "Provisioning of Quality of Service in MANETs performance analysis & comparison (AODV and DSR)," in *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, 2010, pp. V7-243-V7-248.

[10] J. Sen, "A survey on reputation and trust-based systems for wireless communication networks," *arXiv preprint arXiv:1012.2529,* 2010.

[11] J. Li, R. Li, and J. Kato, "Future trust management framework for mobile ad hoc networks," *IEEE Communications Magazine,* vol. 46, pp. 108-114, 2008.

[12] R. Ferdous, V. Muthukkumarasamy, and A. Sattar, "Trust Management Scheme for Mobile Ad-Hoc Networks," in *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*, 2010, pp. 896-901.

[13] I. Ahmad, H. Jabeen, and F. Riaz, "Improved quality of service protocol for real time traffic in manet," *arXiv preprint arXiv:1308.2797,* 2013.

[14] M. Rao and N. Singh, "Quality of service enhancement in MANETs with an efficient routing algorithm," in *Advance Computing Conference (IACC), 2014 IEEE International*, 2014, pp. 381-384.

[15] K. Govindan and P. Mohapatra, "Trust computations and trust dynamics in mobile adhoc networks: a survey," *IEEE Communications Surveys & Tutorials,* vol. 14, pp. 279-298, 2012.

[16] V. Singh and M. Jain, "Secure AODV Routing Protocols Based on Concept of Trust in MANET's," *topology,* vol. 3, 2014.

[17] J. Sen, "A distributed trust management framework for detecting malicious packet dropping nodes in a mobile ad hoc network," *arXiv preprint arXiv:1010.5176,* 2010.

[18] N. Li and S. K. Das, "A trust-based framework for data forwarding in opportunistic networks," *Ad Hoc Networks,* vol. 11, pp. 1497-1509, 2013.

[19] X. M. Zhang, Y. Zhang, F. Yan, and A. V. Vasilakos, "Interference-based topology control algorithm for delay-constrained mobile ad hoc networks," *IEEE Transactions on Mobile Computing,* vol. 14, pp. 742-754, 2015.

[20] R. Kiefhaber, R. Jahr, N. Msadek, and T. Ungerer, "Ranking of direct trust, confidence, and reputation in an abstract system with unreliable components," in *Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC)*, 2013, pp. 388-395.

[21] T. Phakathi, F. Lugayizi, B. Isong and N. Gasela, "Quality of Service of Video Streaming in Vehicular Adhoc Networks: Performance Analysis," *2016 International Conference on Computational Science and Computational Intelligence (CSCI),* Las Vegas, NV, 2016, pp. 886-891.

[22] A. Chen, G. Xu, and Y. Yang, “A Cluster-Based Trust Model for Mobile Ad Hoc Networks,” 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing, 2008

Fine Tune of the Mapping Matrix for Camera Calibration using Particle Swamp Optimization

Tzu-Fan Chen, Wei-Sheng Yang, and Jyh-Horng Jeng
fan1319@gmail.com, jay750216@gmail.com, jjeng@isu.edu.tw
Dept. of Info. Eng., I-Shou University, Taiwan

Abstract—In the traditional camera calibration method, one of the core parameters affecting the calibration quality is the mapping matrices in the *x*- and *y*-direction. In this study, we first calculate the mapping matrices using the traditional method. And then adopt Particle Swamp Optimization (PSO) method to further fine tune the mapping matrices. In the experiments, we use OpenCV packages to perform the original calibration algorithm, and use Python to implement the PSO algorithm to fine tune the mapping matrices. Experimental results show that, we have only minor improvement in terms of re-projection error (since this quantity is coarse), but we do produce better vision effects, especially for the regions distant from the image center.

Keywords: Camera Calibration; Particle Swamp Optimization (PSO); Mapping Matrix; Re-projection Error

Introduction

Camera calibration can be viewed as a pre-processing technique for practical visual applications. For wide-angle and fish-eye lenses, calibration is even more important. It is a key factor of success for the latter applications. Calibration process is indeed nonlinear [1, 2]. To improve the performance of coordinate transformations, Tsai [3, 4] proposed a 2-tage process method, which produces better results but requires high cost computation. Heikkila [5, 6] proposed another 4-stage method to compensate the radial distortion and tangent distortion. The most practical and commonly used method was proposed by Zhang [7] in 2000. His method is widely implemented in Matlab and OpenCV. Recently, there are methods using stereo objects to assist the calibration (3D-calibration) [8]. However, these methods require some specific geometric structures of the objects. The performance is much better than the traditional ones but they are also computation intensive methods.

Particle Swamp Optimization (PSO) is a famous branch of evolutionary computation, which is originated from social psychology principle using the fitness concept. Individuals in the swarm have the ability of communication to exchange information. PSO is first proposed by Kennedy and Eberhart [9, 10] in 1995. The underlying idea is interesting. Wilson [11] found similar behavior from the fish. Reynolds [12] and Heppner [13] found that the birds form some specific figures when flying together to reach a place far away. PSO has man success application such as Queens Puzzle [14], image application [15, 16] and neural networks [17]。

Camera Calibration Model

The most commonly used camera calibration model based on pine-hole mode is depicted in Fig. 1. The transformation model is shown in (1). The matrix $[\mathbf{R}\,|\,T]$, referred to as extrinsic parameter, is a combination of a rotation and a translation from World Coordinate to Camera Coordinate.

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \left[\begin{array}{ccc|c} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{array}\right] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = [\mathbf{R}\,|\,T] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (1)$$

In the Camera Coordinate, we us the property of similar triangles to obtain (2) and (3), where f is the focal length and z_c is the distance of the camera to the world object. For convenient, one adopts homogeneous form as (4). Together with the correction of pixel aspect ratio and center shift of the CCD sensor (c_x, c_y), one re-writes (4) as (5). The matrix A in (5) is called the intrinsic parameter.

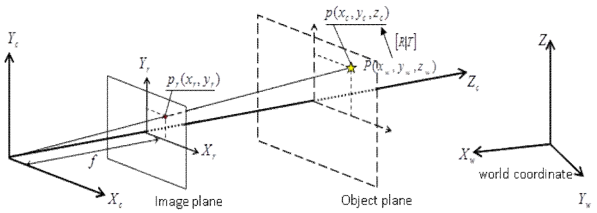


Fig. 1. Pine-hole model for camera calibration.

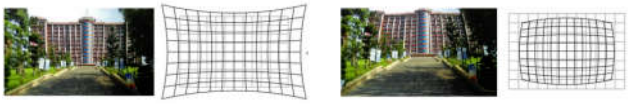
$$\frac{x_r}{f} = \frac{x_c}{z_c} \quad (2)$$

$$\frac{y_r}{f} = \frac{y_c}{z_c} \quad (3)$$

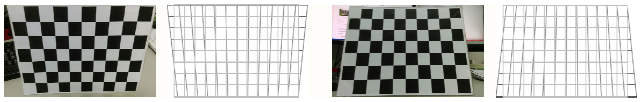
$$S_r \begin{bmatrix} x_r \\ y_r \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} \quad (4)$$

$$S\begin{bmatrix} x_r \\ y_r \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \mathrm{A}\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} \tag{5}$$

There 2 types of distortion induced from the wide-angle and fish-eye lenses, radial distortion and tangent distortion. An example of typical distortion effects is shown in Fig. 2. In the Brown-Conrady model, one merges radial distortion model (6) and tangent distortion model (7) to obtain a homogeneous form (8), in which the nonlinear transform is denoted by Γ .



(a) radial distortion



(b) tangent distortion

Fig. 2. Lens distortions.

$$\begin{aligned} x &= x_r(1+k_1r^2+k_2r^4+k_3r^6) \\ y &= y_r(1+k_1r^2+k_2r^4+k_3r^6) \end{aligned} \tag{6}$$

$$\begin{aligned} x &= x_r + (2p_1x_ry_r + p_2(r^2+2x_r^2)) \\ y &= y_r + (p_1(r^2+2y_r^2) + 2p_2x_ry_r) \end{aligned} \tag{7}$$

$$\begin{bmatrix} x_u \\ y_u \\ z_u \end{bmatrix} = \begin{bmatrix} x_r(1+k_1r^2+k_2r^4+k_3r^6) + (2p_1x_ry_r + p_2(r^2+2x_r{}^2)) \\ y_r(1+k_1r^2+k_2r^4+k_3r^6) + (p_1(r^2+2y_r{}^2) + 2p_2x_ry_r) \\ z_r \end{bmatrix} = \Gamma\begin{bmatrix} x_r \\ y_r \\ z_r \end{bmatrix} \tag{8}$$

Finally, one combines (1), (5), and (8) to form the calibration model as shown in (9). For practical implement, we adopt Zhang's method [7] which utilizes various views of Chessboard to estimate the intrinsic parameters f_x, f_y, c_x, c_y , lens distortion parameters k_1, k_2, p_1, p_2, k_3 , and extrinsic parameters. And then, we can use the intrinsic and lens distortion parameters to perform image calibration.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \Gamma\,\mathbf{A}\,[\mathbf{R}\,|\,T]\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{9}$$

The Mapping Matrix

In the implementation, we adopt the tool in OpenCV-Python packages to calculate the Mapping matrices Mapx and Mapy (abbreviated as Map in together). These 2 matrices provide pixel mapping positions in the x- and y-directions between the true image and the captured (distorted) one. They can be regarded as a simulation of a camera which distorts an image in the capture process. Therefore, whenever these 2 matrices are obtained, the correction process can be done by a process called Remap which perform inverse mapping together with image interpolation. The relation between Map and Remap is illustrated in Fig. 3.

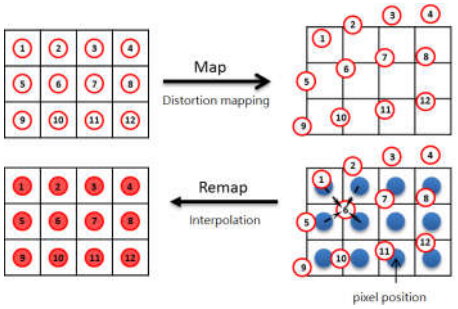


Fig. 3. Map and Remap with interpolation

Here since the originally obtained matrices Mapx and Mapy are not precise, especially in the regions distant from the image center, we use PSO to further fine tune them. The effect of Map is shown in Fig. 4. The left image is an ideal box image, while the right one is the transformed (distorted) one. As observed, distant regions have bigger distortions.

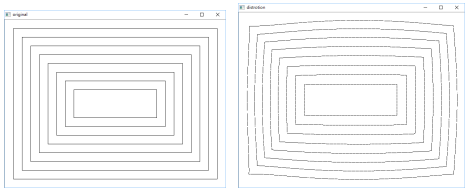


Fig. 4. Direct mapping of Mapx and Mapy on as ideal image.

Particle Swarm Optimization, PSO

The simplest model of PSO is given in (10) and (11). For each particle, let x be the particle position and V be the particle velocity. Let P be the best position of the particle x , i.e., Pbest and G be the best position of the swarm, Gbest. The subscription i represents i -th iteration.

$$V_{i+1} = \omega \cdot V_i + Cp \cdot Rp \cdot (P_i - x_i) + C_G \cdot R_G \cdot (G_i - x_i) \tag{10}$$

$$x_{i+1} = x_i + V_{i+1} \tag{11}$$

One of the main reason rendering PSO useful is the coefficients of the 3 terms in (10). The coefficient ω stands for the inertial term. The coefficients C_P and C_G are random numbers between 0 and 1, and different for each component.

Experimental Results

In this study, we adopt Python (2.7.5), numpy (1.9.2) and OpenCv (cv2 3.0.0) as our software platform. Testing images are of size 480×640. A set of chessboard are captured in various aspect. Some examples of images are shown in Fig. 5. After the calibration parameters are obtained using the training set, some images from the testing set are tested. The re-projection error can be calculated on the chessboard corners detected previously. It is the averaged difference of calibrated and then distorted one over the training set.

Fig. 5. Examples of captured chessboard images.

The results of fine tune using PSO are listed in Table 1-3. Table 1 shows the result using various number of particles in PSO. We observe that there are reductions of re-projection error using PSO, although the amount is not significant. Table 2 shows the results using various number of iterations and Table 3 shows those using various parameter combinations.

Table 1. Test on various number of particles.

Parameter settings			
epoch =3, generation=100, ω=0.6, C_P=1.7, C_G=0.6			
No.	particles	Re-projection Error	Error Difference
1	100	2.75207381	- 0.05944
2	200	2.75140738	- 0.06011
3	300	2.75148955	- 0.06003

Table 2. Test on various number of iterations.

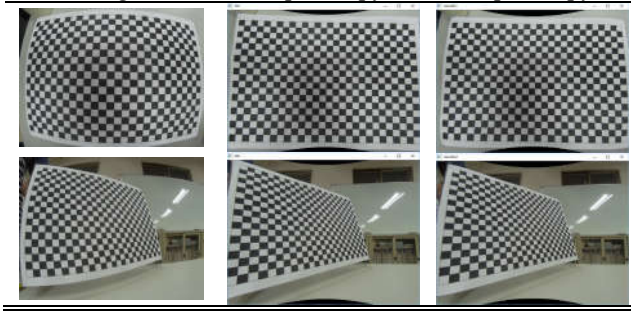
PSO settings			
epoch=3, particles=100, ω=0.6, C_P=1.7, C_G=0.6			
No.	iterations	Re-projection Error	Error Difference
1	100	2.77927487	- 0.03224
2	200	2.77912801	- 0.03239
3	300	2.77897274	- 0.03255
4	400	2.77920540	- 0.03231

Table 3. Test on various parameter combinations.

基本參數設定					
epoch = 5, 粒子數=10, generation=100					
No	ω	C_P	C_G	Re-projection Error	Error Difference
1	0.6	1.7	1.7	2.75143589	- 0.06008
2	0.6	0.6	1.7	2.75171042	- 0.05981
3	0.6	1.7	0.6	2.75132992	- 0.06019
4	0.6	0.6	0.6	2.75160842	- 0.05991
5	0.6	0.1	1.7	2.75167209	- 0.05984
6	0.6	1.7	0.1	2.75165152	- 0.05986
7	0.6	1.2	0.6	2.75167752	- 0.05984

The visual results are illustrated in Table 4. In the left column, there are 2 captured images. The middle column shows the results of the originally calibrated images and the right column shows those using fine-turned parameters. Compare the left corner of the images in the bottom middle and bottom right. There are still apparent distortions in the former, while the latter exhibits better visual effect.

Table 4. Comparison of visual effects.

Original Image	Original Mapx, Mapy	Fine-tuned Mapx, Mapy
		

Summary

In the field of camera calibration, the distortion model can be represented using only few parameters. That is, it is impossible to adjust those parameters to obtain better results because of the tight relaxation property of the model. Instead, we perform the fine tune directly on the mapping matrices of the distortion model, which consist of thousands of parameters. Therefore this method is somewhat too much relaxed. In the future, we will try to impose some constraints in order to reduce the amount of parameters.

Reference

[1] W. Faig, “Calibration of close-range photogrammetric systems: Mathematical formulation,” Photogramm. Eng. Remote Sens., vol. 41, no. 12, 1975..

[2] M. A. Penna, “Determining camera parameters from the perspective projection of a quadrilateral,” Pattern Recognit., vol. 24, no. 6, pp. 533–541, 1991.

[3] R. Tsai, “A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses,” IEEE J. Robot. Autom., vol. 3, no. 4, pp. 323–344, 1987.

[4] R. Tsai, “An efficient and accurate camera calibration technique for 3D machine vision,” in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 1986.
[5] J. Heikkila and O. Silvén, “A four-step camera calibration procedure with implicit image correction,” in Computer Vision and Pattern Recognition, 1997.
[6] J. Heikkila, “Geometric camera calibration using circular control points,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 10, pp. 1066–1077, 2000.
[7] Z. Zhang, “A flexible new technique for camera calibration,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 11, pp. 1330–1334, 2000.
[8] C. Chen, C. Yu, and Y. Hung, “New calibration-free approach for augmented reality based on parameterized cuboid structure,” in Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, 1999, vol. 1, pp. 30–37.
[9] J. Kennedy, “Particle swarm optimization,” in Encyclopedia of machine learning, Springer, 2011, pp. 760–766.
[10] R. Eberhart and J. Kennedy, “A new optimizer using particle swarm theory,” in Proceedings of the sixth international symposium on micro machine and human science, 1995, vol. 1, pp. 39–43.
[11] E. O. Wilson, “Sociobiology, the new synthesis Belknap Press,” Camb. Mass, 1975.
[12] C. W. Reynolds, “Flocks, herds and schools: A distributed behavioral model,” ACM SIGGRAPH Comput. Graph., vol. 21, no. 4, pp. 25–34, 1987.
[13] F. Heppner and U. Grenander, “A stochastic nonlinear model for coordinated bird flocks,” Ubiquity Chaos, pp. 233–238, 1990.
[14] X. Hu, R. C. Eberhart, and Y. Shi, “Swarm intelligence for permutation optimization: a case study of n-queens problem,” in Swarm intelligence symposium, 2003. SIS’03. Proceedings of the 2003 IEEE, 2003, pp. 243–246.
[15] K. Deep, M. Arya, M. Thakur, and B. Raman, “Stereo camera calibration using particle swarm optimization,” Appl. Artif. Intell., vol. 27, no. 7, pp. 618–634, 2013.
[16] A. M. Rather, A. Agarwal, and V. N. Sastry, “Recurrent neural network and a hybrid model for prediction of stock returns,” Expert Syst. Appl., vol. 42, no. 6, pp. 3234–3241, 2015.
[17] P. Tiwari, S. Ghosh, and R. Sinha, “Classification of two class motor imagery tasks using hybrid GA-PSO based k-means clustering,” Comput. Intell. Neurosci., vol. 2015, no. 59, 2015.

Analysis of Modeling Design of Control Java Programming Exercise for Game Subjects and a Traveling Machine with LEGO Mindstorms

Shibo Ryu, Kento Tsuji and Hiroyuki Tominaga
Kagawa University
2217-20, Hayashi-machi, Takamatsu-shi, Kagawa, Japan
s14g486@stmail.eng.Kagawa-u.ac.jp, s12t241@stmail.eng.Kagawa-u.ac.jp,
tominaga@eng.Kagawa-u.ac.jp

ABSTRACT

We have proposed a control system programing exercise using Traveling body equipped with LEGO Mindstorms microcomputer. It is practicing in classes of our information engineering colleges. The execution environment is leJOS of Linux base, and development environment is Object-Oriented Java language. As a group learning of the project unit, tackle with game project including technical elements. In this paper, we are focus on modeling design of applied Problems which is UML diagram described on students report. We are compare and discuss the item should be described on each diagram with actual answer examples.

KEYWORDS

Control system programing; Java exercise; Modeling design; UML diagram;

1 INTRODUCTION

1.1 EV3 kit of LEGO Mindstorms

LEGO Mindstorms is very useful educational toys. It is developed by LEGO Company and MIT [1]. It is a simple robotics kit for robot construction and control programming. The kit contains EV3 micro-computer, motors, various sensors, many mechanical parts and traditional bricks. EV3 software is prepared as the standard environment for visual programming. It has user-friendly interface like brick construction and flowchart drawing. A user makes a control program on PC and transports it to EV3 by a USB cable or wireless of Bluetooth. The robot moves autonomously without PC. It detects the outer state by sensors and moves as the reaction and response according to the program. The kit is utilized in various educational scenes for engineering practices and contest events.

1.2 Educational application of LEGO Mindstorms

We have proposed an introductory programming exercise using LEGO Mindstorms robot kit [2][3][4]. The target learners are mainly very beginners of entrant students in engineering college. It is as a pre-education practice before the proper programming lesson. It offers the first experience opportunity of simple software development with group collaboration. We expect to make them feel so-called craftsmanship sense to promote their motivation as engineer. Our educational purpose is to raise the problem solving skill by programming as a solution tool. Controlling a concrete and physical object like a LEGO robot by a program must give students strong impressions about a sense of accomplishment.

2 CONTROL PROGRAMMING FOR EMBEDDED SYSTEM

2.1 Education of control programing of embedded system

Recently, the importance of control technique in embedded system has been increasing. Though it is necessary as a field of information engineering education, the educational cost is rather high. The feature of embedded system is by using device like motors and sensors to realize relation with the real world, and solve various problems arising in the interface with the outside world such as noise and error.

2.2 The educational purpose of exercise

In the information department college of our university, since 2010 we are carrying out control system programing exercise by simple robot and game projects in compulsory subject "Experiment 2" of specialized course.
In the exercise curriculum, as various development method like Java grammatical matter, introduce object-oriented programing, Foundations of Software Engineering will be after UML notation learning. The educational purpose of exercise, there are control system programing as the first step of embedded system, modeling design along the Object-Oriented, trial test aware of reality environment and physical constraints, and projects in group.

2.3 Modeling design and UML diagram

So far, we have been analyzed evaluation of student's questionnaire [3], description contents of report [4] in 2015 practice. Also, in 2016, we attempted to analyze answer programs in applied problems with code indicators [5]. In this paper, we are analyze Modeling design which was described in report with quantity and contents of UML diagram, and it will lead to more precise evaluation.

2.4 Development environment LeJOS for EV3 kit

In our exercise from 2010 to 2014, we adopted ROBOTC environment as extended C language with pseud multi-task processing for old version NXT of LEGO Mindstorms. From 2015, we changed leJOS environment, which is a Linux OS with Java virtual machine by Oracle. It contained Java Runtime library for new version EV3 as embedded system. You can use Standard Java API as object-oriented programming. It is available for parallel processing with multi-thread.
In addition, we offer original library for sensor control. The color sensor consists of both bodies of light emission and light acceptance. When the light emission is off, it measures RGB values of outer space. When the light emission is on, it measures the values of reflection from neighboring sphere. In our libraries, the color sensor has a converting method of HSV values with bicone model. The value V can be used for measurement of lightness as a light sensor. The value H can be used for color pattern distinction, which returns a typical color name. It also has a calibration method to decide some threshold values.

3 ABOUT OUR EXERCISE

3.1 Outline of our exercise

In this exercise, we are prepared the prescribed robot suitable for the traveling body for line trace (Fig.1). There are 4～5 students in 1 group, then give them 2 regulated robots and make the group into 2 units. For the problems, students should control robot to finish traveling on the course and area of the game filed, and realize mission element in game sense. In the group, members should take charges for design of strategy, implementing the program, and verify the operation. The points of performance certification was total by traveling point by running time and mission point by achievement level.

3.2 Lesson project and game projects

In this exercise, Game projects also was been reconstructed with the new kit. Project group was been organized in Table 1 [6][7]. For each project, we prepared example program showing usage of sensors, several basic problem corresponding to individual technical items, comprehensive applied problems with performance certification. Applied problems are game projects using dedicated game field. During the projects, we presented the scoring rule of performance certification, and set up intermediate goal for each mission element. For students, after learning control by the basic problem, examine the rules of applied problems, select the mission elements, and consider the strategy. At the final contest of the synthesis problem, we will tackle about three

weeks [8]. In the past, there was a problem which is transport control by arm mechanism. But the migration of mechanism is insufficient in new version of regulated robot, so the problem has been omitted.

3.3 Support tools for our exercise

To support this exercise, we are building Class content management site which is LegoWiki based on PukiWiki and LegoPress based on WordPress. The purpose of the project is posting projects scoring rule, technical item like characteristics of sensor, program of example project, and referee sheet of performance certification.

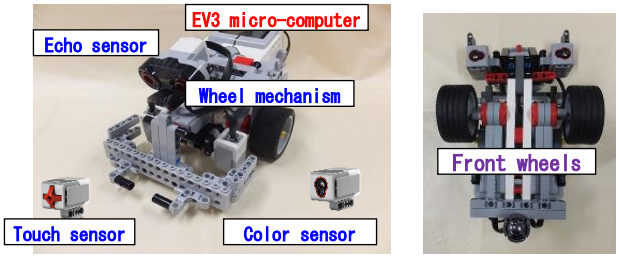


Figure 1 Traveling machine LeBot of our exercise

Table 1 Curriculum

Week	Project	Game subject
1	Event driven and task management	10 Basic problems
2	Running control by wheel mechanism	21 Figure Following
3	neighbor detection by light sensor	30 Line Trace 31 Performance Driving 32 Area Sweeping
4	Faraway detection by echo sensor	41 Target Access 42 Target Round
5	Gradation distinction by color sensor	51 Garage Parking
6 7 8	Great race contest as final project	01 Outside Running 02 Inside Running

4 CONTENTS OF APPLIED PROBLEM

4.1 Lesson 1

In the Lesson 1, a student learns basic control with API for LEGO Mindstorms. Student will learn about event driven by Touch sensor or button, parallel processing by multi-tasking, use of flag or status, and implementation of event queue. In this lesson, student will use sensor-related libraries which has been prepared. Because of the answer program is just following the example program, so it is not worth be the subject to analysis or evaluation.

4.2 Lesson 2

In the Lesson 2, a student learns running control of driving and steering by wheel mechanism with independent left-and-right wheels. In the game subject 21 "Figure Following", he must realize running the black line course with sequential control. The course has two partial areas in the game field (Fig.2). The first half seems L shape as straight course. It requires a simple program with only parameter setting for straight run and spin turn. The second half seems "3" shape with two half circles. It requires a program with rather careful parameter setting for curving run. In modeling design, he must consider several state transition according to the partial course. The triggers are by time or distance.

4.3 Lesson 3

In the Lesson 3, you learn neighboring detection by color sensor. In the game subject 30 "Line Trace" (Fig.3), they must realize line trace running with feedback control. LeBot machine runs on the center of black lines by power control of both motors according to each light sensor.
Moreover, in the game subject 31 "Performance Driving" (Fig.3), LeBot must also perform some given missions by detecting color tiles like "Dance Action". It does spin turn at red tile and makes beep sounds at each green tile. The score of the game is the sum of running point by the speed and mission point by achievement degree. Starting operation is by button press. Stopping operation is by conflict to tower object with touch sensor. In the game subject 32 "Area Sweeping" (Fig.4), LeBot with bumper attachment must keep in the area rounded by the black line. It must sweep out some balls as obstacle by zig-zag moving for all area like "Roomba Cleaner".

4.4 Lesson 4

ISBN: 978-1-941968-42-0 ©2017 SDIWC

In the Lesson 4, you learn faraway detection by radar attachment with echo sensor. The sensor can measure a distance from a target object by ultrasonic echo location. However, there are some bothersome features about directivity in fan shape, delay by sampling interval and measurement error.
In the game subject 41 "Target Access" (Fig.5 left), LeBot searches a tower object during spiral movement. The attachment is placed in the front top. When the tower is found, it approaches it as close as possible without touch like "Chicken Run". In the game subject 41 "Target Round" (Fig.5 right), LeBot must go around a tower object by without touch like "Swing By". The attachment is placed in the left side top. It must also perform driving of course out and course back as line trace.

4.5 Lesson 5

In the Lesson 5, you learn gradation distinction of gray scale by color sensor. In the game subject 51 "Garage Parking" (Fig.6), Lebot must perform garage in and out on a sheet. It must run across the gradation by a difference value of both color sensors. It also must turn in light gray band and go to a blue line and back.

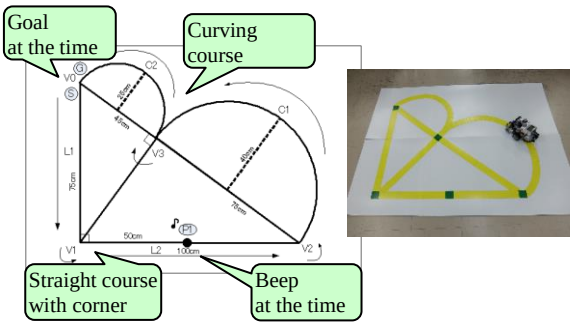


Figure 2 Game "Figure Following"

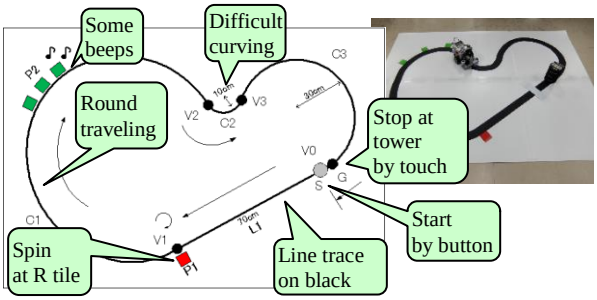


Figure 3 Game "Performance Driving"

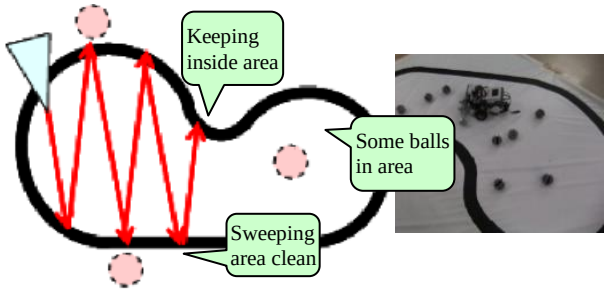


Figure 4 Game "Area Sweeping"

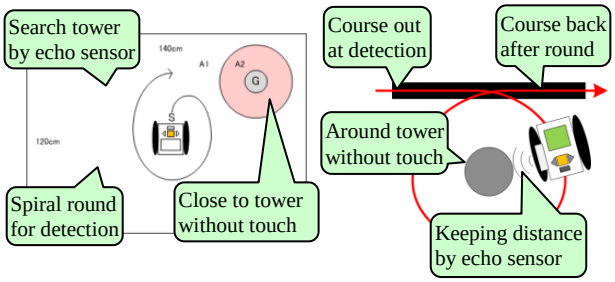


Figure 5 Game "Target Access" and "Target Round"

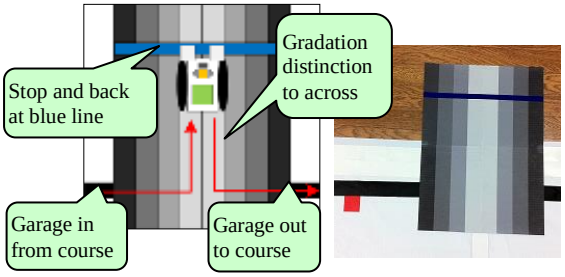


Figure 6 Game "Garage Parking"

5 UML DIAGRAMS IN SUMMARY REPORT

5.1 Advance Explanation for 2015 and 2016

During the 2015 reports, describe outline design with diagram was been ordered. But mostly of student were only describe wit flow chart. In 2016, in advance, we explained area division of game field to several areas according to the characteristic, and consider the optimal behavior in each area. This is to emphasize the necessity of State machine diagram.

5.2 UML Diagrams in summary report

During the 2016 reports, 45 UML diagrams were described in total of 10 groups for 6

ISBN: 978-1-941968-42-0 ©2017 SDIWC

problems. There are 7 class diagram, 32 state machine diagram, 2 activity diagram, and 4 Sequence Diagram. For drawing UML diagram, few tool like Microsoft Visio, astah* community [11] was been used. There were also few of groups drawing diagram by Free-hand. We focused quantity of components of UML diagram in summary reports. For each UML diagram, we enumerated frequency number of nodes and arcs as relation between nodes. Table 3 it shows the average values.

Table 2 Frequency of UML Diagrams in reports

Diagram	**21**	**31**	**32**	**41**	**42**	**51**	Total
Class	0	1	1	1	2	2	7
State Machine	7	8	3	3	8	8	37
Activity	0	0	0	0	2	0	2
Sequence	2	1	1	0	0	1	5

Table 3 Quantity average of UML Diagrams in reports

Average	Elem	**21**	**31**	**32**	**41**	**42**	**51**
Class	Node	-	12.0	13.0	9.0	8.5	8.5
	Arrow	-	10.0	7.0	6.0	7.5	6.0
State Machine	Node	11.1	8.1	9.3	10.0	9.9	12.0
	Arrow	11.0	9.5	12.0	11.3	10.9	13.4
Activity	Node	-	-	-	-	26.5	-
	Arrow	-	-	-	-	33.0	-
Sequence	Node	5.0	4.0	6.0	-	-	7.0
	Arrow	14.0	4.0	8.0	-	-	9.0

6 CLASS DIAGRAM AND OBJECT DIAGRAM

6.1 Outline of Class Diagram and Object Diagram

Class Diagram as structural diagram enumerates all classes in the system and describes relationship between classes like E-R Diagram (Entry-Relationship). Object Diagram describes relationship between instances. The traveling machine must be structured by detection mechanism and wheel mechanism. These mechanism are structured in hierarchical by several components, such as, various sensors and stepping motors. Necessary classes in given execution library are also described (Fig.7). A light sensor class in Class Diagram represents available function to detect lightness of outer space or neighboring sphere. Both instance objects of light sensor in Object Class show different roles by left and right position.

6.2 The exemplar of Class Diagram

As the Game Subject 31 "Performance Driving", there are Input class and Output class. Input class as detection mechanism includes LightSensor, ColorSensor and TouchSensor as sub-class. Output class as operation mechanism includes Motor and Sound as sub-class. Moreover, Mission class is needed which has some Input and Output classes according to each mission performance. It includes LineTrace, ButtonStart and GoalStop as sub-class, which use LigntSensor. It also includes SpinTurn and AlertBeep as sub-class, which use ColorSensor. LineTrace and AlertBeep class has Runnable interface for multi-thread process. If they are single thread, beeping process and driving process don't work in parallel execution. The Main class consists of have these Mission classes. (Fig.7)

6.3 Consideration about answers of Class Diagram

In each game subject, many of group mentioned Class Diagram in summary reports. The description quantity of nodes and arrows in all diagrams was sufficient for the detail. Some description errors were founded because of inexperience of UML editors. Object Diagram was few mentioned. Many group may regard Sequence Diagram as substitution of Object Diagram.

7 STATE MACHINE DIAGRAM

7.1 Outline of State Machine Diagram

State Machine Diagram as behavior diagram defines some state and describes transition between states. It is important to describe an explicit trigger event with condition of state transition. Students often confuse state transition and a flowchart. While a node of flowchart is a process by changing another node after finishing, a state node in the diagram

represents a situation with an interval which is still at the same node until a trigger event.
Initial and end state must be appeared clearly in a diagram. You may use Composite State including some sub-state as nest structure. As game subject 21 "Figure Following", there are simple state of straight, spin and rotation. Triggers of transition occur by timer event or rotation count event.

7.2 The exemplar of State Machine Diagram

As State Machine Diagram, we introduce the exemplar of game subject 51 "Garage Parking" in Fig.7 (b). It has 6 composite states according to the position in course or the sheet. The triggers of transition are detection of color change. It may add time out event. They may help to keep running though giving up the mission in the middle.
State (1): Running along the course
State (2): Searching the entrance position into the sheet
State (3): Driving to the stop line with forward
State (4): Driving from the stop line with backward
State (5): Searching the exit position out of the sheet
State (6): Running along the course

7.3 Answer examples of State Machine Diagram

We discuss some answer examples of State Machine Diagram about "Garage Parking" (Fig.8). (b) (d) only have 1 composite state so the granularity is inappropriate. Especially, (d) has no sub-state, it is unknown what Kind of behavior has be done. In (a), the state name is the action itself. Therefore, internal actions of state are not been described. In (c), garage parking after stop on blue line state, retreat within the garage sheet state was not been described, back to black line trace state soon.

7.4 Consideration about answers of State Machine Diagram

Because we emphasized importance of the diagram for game projects in lesson, almost group drew them. If you focus on high speed running or mission accomplishment without considering the internal structure of regulated robot, state machine diagram is very important. The diagram seems looks like flowchart, therefore some of the student draw state machine diagram in that sense. About the description error, there are Simple symbol mistake or no arrow of transition. Also, there are some mistake like internal action or transition conditions are didn't described.
The quantity of state machine diagram, which was small overall. The reason that the number of node is because composite states and sub-states are not properly set. Before drawing state machine diagram, it is necessary to divide the behavior to capture project into several states. Inside of composite state, initial state and final state should be described. Also, by using parallel state to represented parallel processing correctly, the number of nodes will also increase.

8 ACTIVITY DIAGRAM AND SEQUENCE DIAGRAM

8.1 Outline of Activity Diagram

Activity Diagram describes the execution order of several processes like a flowchart. It can deal synchronization of parallel processing by symbol "sync bar". Moreover, you can use nest structure for complex process according to the granularity of tasks. In our game subjects, description of behaviors from start to goal by a traveling machines needed. Only in the game subject 42 "Target Round", some of groups has been describe activity diagram. For the description error, almost the use of symbol is incorrect. For example, the symbol of end node and final control was been confused. About the quantity of diagrams, it was appropriate.

8.2 The Exemplar of Activity Diagram

As activity diagram, the overall flow that point out robot's behavior is necessary. In the game project "Target Access", branching conditions are important. When echo sensor or touch sensor detected the target, the stop processing

should be activated. In the target object has been detected and access processing, using fork node and join node to describe parallel processing of forward running and distance measure.

8.3 Outline of Sequence Diagram

Sequence Diagram as interactive diagram describes message passing between objects according to time series. In Sequence Diagram, to express the control structure like branch, loop, parallel processing, using combined fragment is necessary.
There are many of quantity should be descried in sequence diagram, therefore the difficulty is slightly higher. So there were little description in the reports. Description error are due to insufficient understanding of grammar. For example, confused with name of message and lifeline, didn't describe execution specification, inappropriate to distinguish messages between synchronous and asynchronous. The number of nodes was appropriate, but the message could not be distinguished, so required arcs was insufficient.

8.4 The exemplar of Sequence diagram

When draw sequence diagram, you have to understand message passing between classes which are described on class diagram. In the game projects "Performance Running", express the method processing between mission class and control class is very important. When executing Line Trace, it is important to express paralleling processing of traveling and sounding beep by using fragment. Also, for the color taken by both light sensors at Line Trace, express conditional branch during selection between forward running and spin turn by fragment. When sounding by detected the green, using fragment to express iteration processing.

9 CONCLUSION

We proposed an applied group exercise of control Java programming for a travelling machine with LEGO Mindstorms. The exercise is for introductory object-oriented programming of simple embedded system control. It is also for the first training of object-oriented development by UML modeling design. We offered some game subjects in each project with some technical items according to the lesson curriculum.
It is performed in require subject "Experiment 2" as required subject in information engineering college. It is held in quarter of 8 weeks. We discussed answers of UML diagram in summary reports in 2016. We analyzed the description quantity and compared them with the exemplar. Generally, student get strong impact from flowchart. For modeling design of object-orientation, it was been confirmed that practical understanding is insufficient. In future work, presentation of exemplar UML diagram, Create check sheet, publish evaluation measure.

REFERENCES

[1] LEGO Company, LEGO. com Mindstorms Home, http://mindstorms. lego. com/eng/default. asp.

[2] Y. Onishi, H. Tominaga, T. Yamasaki, "GoalPost: LEGO Programming Exercise Support", Proceedings of AMT 2005, pp.305-306, 2005.

[3] Y. Onishi, H. Tominaga, T. Hayashi, T. Yamasaki, "Exercise Analysis and Lesson Plan with Robot Behavior in LEGO Programming Contest for Problem Solving Learning", Proceedings of ED-MEDIA 2006, pp.1943-1950, 2006.

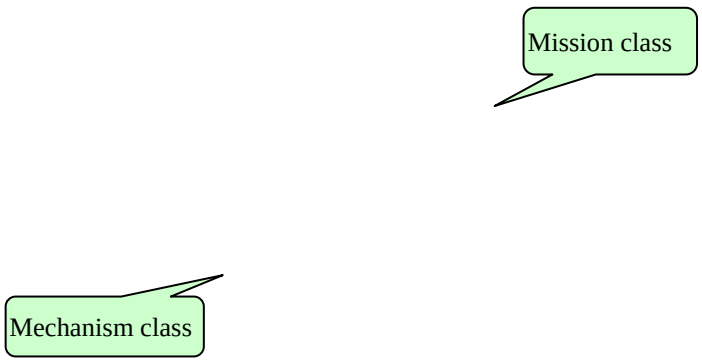
[4] Y. Onishi, H. Tominaga, T. Hayashi, T. Yamasaki, "GoalPost: LEGO Programming Exercise Support for Problem Solving Learning with Strategy Design Tool", Posters of ICCE 2006, pp.53-56, 2006.

[5] H. Tominaga, Y. Onishi, T. Hayashi, T. Yamasaki, "LEGO Robot Programming Exercise Support for Problem Solving Learning with Game Strategy Planning Tools", Proceedings of DIGITEL 2007, pp. 81-88, 2007.

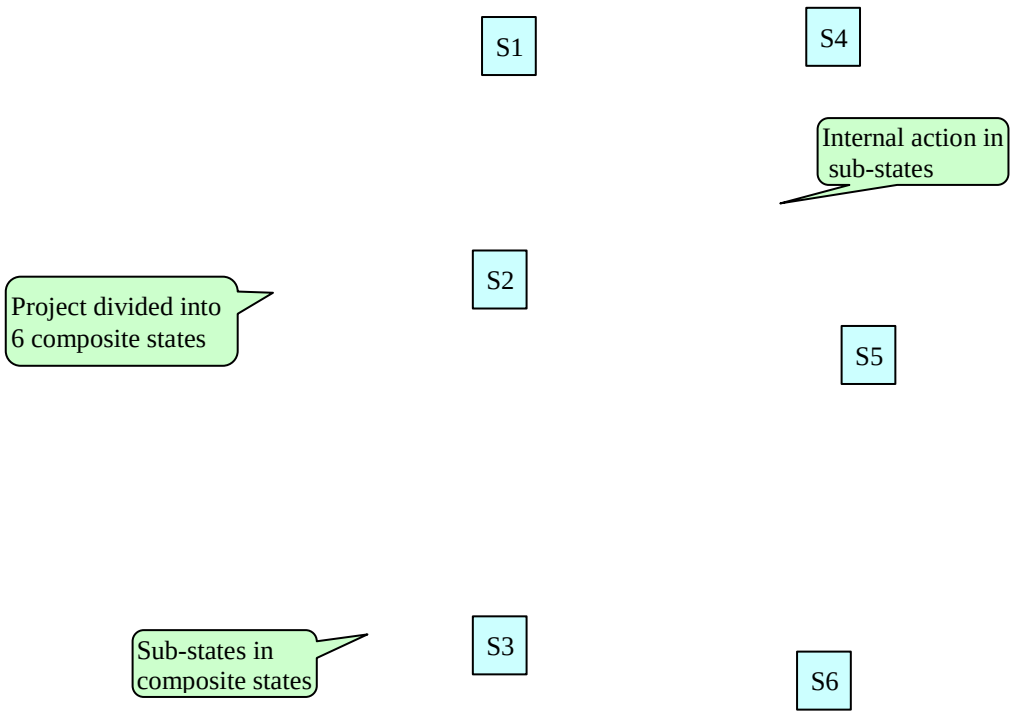
[6] S. Kato, H. Tominaga, "A practical Lesson of Introductory Programming Exercises with LEGO Robot Control and Game projects", Proceedings of ED-MEIDA 2009, pp.1747-1752, 2009.

[7] S. Kato, H. Tominaga, "A Style and Tool for Group Exercise of Introductory Programming with LEGO Robot Control as Pre-Education Event", Proceedings of ITHET 2010, pp.259-267, 2010.

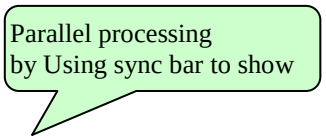
[8] S. Kim, J. Jeon, "Introduction for Freshmen to Embedded Systems Using LEGO Mindstorms", IEEE Education Society 2009, pp.99-108, 2009.



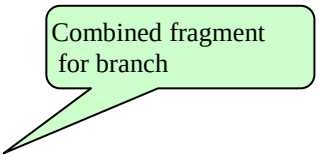
(a) Class Diagram of "Performance Driving"



(b) State Machine Diagram of "Garage Parking"



(c) Activity Diagram of "Garage Parking"



(d) Sequence Diagram of "Performance Driving"

Figure 7 Exemplar samples of UML diagrams

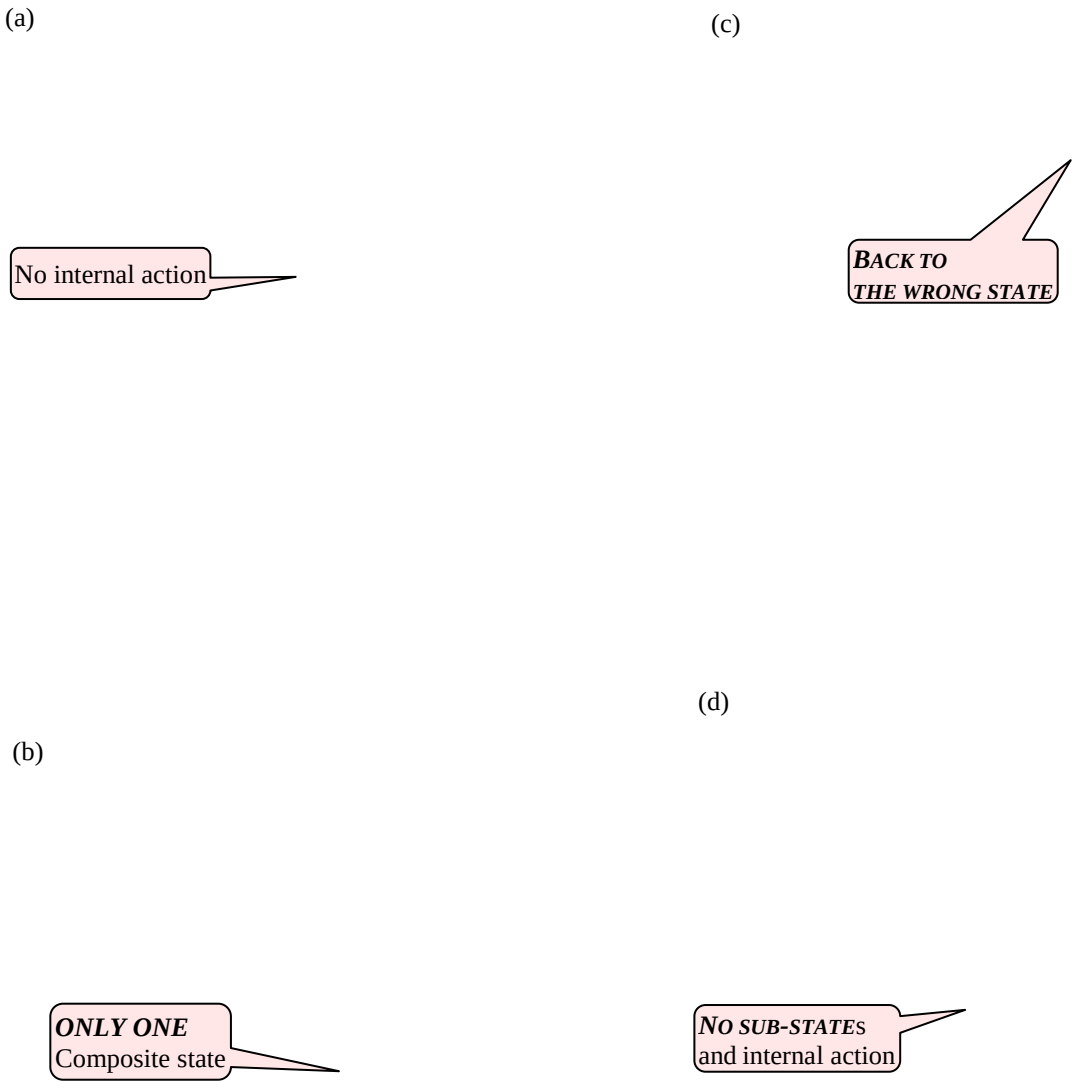


Figure 8 Answer examples of State Machine Diagram

A Prototype System to Browse Web News using Maps for NIE in Elementary Schools in Japan

Yutaka Uchiyama [*1] Akifumi Kuroda [*2] Kazuaki Ando [*3]
[*1, 2] Graduate School of Engineering, [*3] Faculty of Engineering
Kagawa University
Takamatsu, Kagawa, JAPAN
[*1] s17g453@stu.kagawa-u.ac.jp, [*3] ando@eng.kagawa-u.ac.jp

ABSTRACT

In many elementary schools around the world, NIE (Newspaper In Education) that uses newspapers as teaching tools has been implemented. However, the contents of newspaper articles are difficult for elementary school children. It is also not easy for children to find interesting articles from the newspapers. In this study, we propose a system to browse Web news using maps, in order to provide support for NIE in elementary schools in Japan. Children can find, confirm and learn various events which occurred near their own town by using the proposed system. This paper proposes an interface and support functions of a prototype system, and explains results of preliminary evaluations of the system. Finally, we improve a part of the interface of the proposed system based on the evaluation results.

KEYWORDS

Newspaper in Education, Interface using Maps, Browsing Web News, Elementary School Children

1 INTRODUCTION

In recent years, NIE (Newspaper in Education) has been implemented in educational institutions including elementary schools around the world. Newspaper articles are used as teaching materials for NIE. In the practical report for NIE by the Japan Newspaper Publishers & Editors Association [1], it is reported that NIE can grow elementary school children's interest in society and improve their reading comprehension. It is also effective to develop communication skills, because opportunities of conversation related to the contents of news increases between parents and children.

Paper newspapers are generally used in NIE. In recent years, the use of Web news is receiving attention in NIE, because newspaper publishers provide newspaper articles as Web news on their Web sites. However, these articles are not written for children. Most of elementary school children rarely understand the contents of the articles, because the articles contain difficult words and expressions for children. Therefore, it is not easy for children to find interesting articles [2]. In order to improve this situation in NIE, we consider that a support system to choose and browse Web news articles is necessary.

On NIE in elementary schools in Japan, teachers take up news articles related to children's own town frequently. The children learn various topics, features and events of their region, and find interesting articles in their region and elsewhere. Therefore, information about locations is one of important factors for NIE in elementary schools. Furthermore, the children start to learn a map in the curriculum for third or fourth grade of elementary schools in Japan.

From the above points, we propose a support system to browse Web news using maps in this research. By displaying news articles on the maps, the children can choose and browse interesting articles from the maps, and understand the relationship between the contents of the article and the location easily.

This paper describes the outline of the prototype system to browse Web news using maps for NIE in elementary schools in Japan, and the results of preliminary evaluations.

2 RELATED WORK

There are a few systems to browse Web news using maps, such as Newsstand [3], [4] for English speakers and Mapnews [4] for Japanese speakers. Figure 1 shows a screenshot of the interface of Newsstand. These systems extract information about the location from each Web news article and place a marker indicating the article on the maps based on the extracted information. The system using maps as a browsing interface is very useful for finding, choosing, and browsing articles related to geographic locations that readers are interested in, because they can visually recognize the relationship between the contents of the articles and the locations.

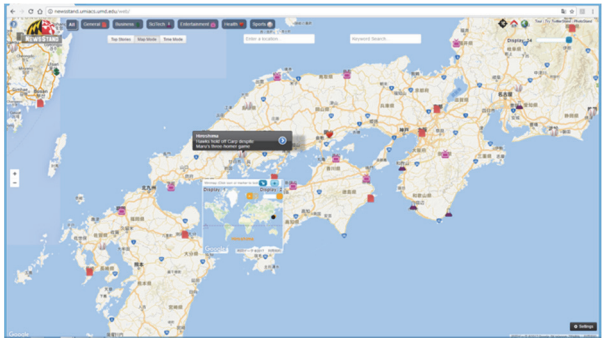


Figure 1. A screenshot of the interface of Newsstand [4]

However, when these systems are used for NIE in elementary schools, there are the following problems:

(1) These systems display unnecessary news articles on NIE in elementary schools, because these systems do not have a function to filter out the articles;
(2) Children cannot use categories of each article for choosing articles, because Mapnews does not classify markers on the maps by categories;
(3) Children cannot narrow down articles by some conditions such as categories, the date of issue, a period of time;
(4) It is difficult to grasp the outline of an article from information displayed, because only the title or headline of the article is displayed after clicking a marker;
(5) Children cannot understand the contents of news articles because the articles contain difficult terms and phrases;

Therefore, it is difficult to use these systems for NIE in elementary schools in Japan.

3 PROPOSED SYSTEM

3.1 Basic Functions

In order to clear the problems (1) to (5), we propose a system to browse Web news using maps for NIE in elementary schools in Japan. The proposed system has the following functions:

(1) Filtering unnecessary articles for NIE
(2) Color coding of markers based on categories
(3) Narrowing down articles
(4) Simplification of articles

Figure 2 shows a screenshot of the main interface of our prototype system with above functions (1) to (3).

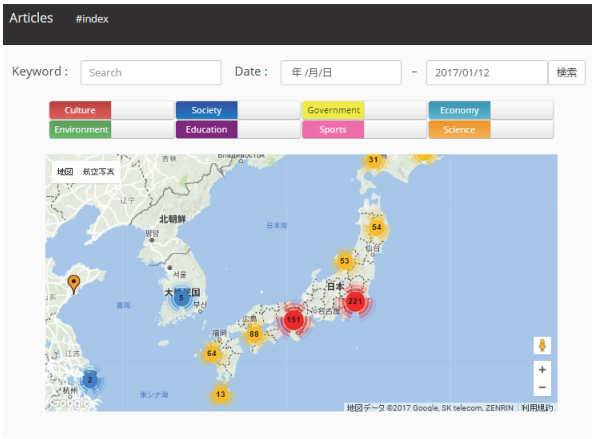


Figure 2. A screenshot of the main interface of the proposed system.

As shown in Figure 2, the map based on the Google Maps API [5] is displayed at the center of the main interface. The proposed system extracts information about the location from each Web news article, and place markers indicating the articles at the appropriate positions on the map. The neighboring markers are consolidated based on zoom levels, and shown the number of the consolidated articles

is displayed in the circles. As shown in Figure 3, each marker is given different color according to the category of each article. The input form to narrow down articles by keywords, a period of time, and categories is arranged on the upper side of the map. If a marker is clicked, an information window consisting of the title, image, date of issue, publishing company, hyperlinks to all locations of the article is displayed.

Below the map, the list of articles placed at the same location is displayed. All locations in the text of each article are linked to maps. Readers can confirm the locations on the maps by clicking the interesting locations in the text or the information window.

Figure 3. A screenshot of an information window and a list of articles placed at the same location

3.2 How to Place Markers Indicating News Articles

The process of placing markers indicating news articles consists of the following three steps.

Step 1: Collection of Web news articles and information extraction from the articles

This system collects news articles from web sites of various newspapers. The articles such as drugs and homicides are filtered out from the set of articles, because they are rarely used for NIE in elementary schools. Then, the title, body, image, URL, date of issue, publishing company are extracted from the collected articles.

Step 2: Extraction of locations from the articles

The system extracts information about locations from each article by using Japanese Named Entity Extraction API by goo Lab [6]. By our preliminary experiments [7], we have confirmed that approximately 83.5% of the news articles contain information about locations.

Step 3: Placing markers on the maps

The system gets latitude and longitude of all locations of the article by Google Maps Geocoding APIs. In the news text, a location is abbreviated after the first mention. If the same coordinates are obtained by APIs, the longest name of locations is selected as one of the representative locations of the article. If the coordinates of "Kagawa" and "Kagawa Prefecture" are same, the latter is selected.

The proposed system places markers indicating the articles at the appropriate position on the maps based on the coordinates. If there are multiple articles in the same locations, the marker of the latest article is set as the representative one.

3.3 Function of Narrowing Down Articles

In order to find interesting articles, the proposed system has the function of narrowing down articles by categories, a period of time, and keywords. Keywords and a period of time are input from the input form. A period of time is set to a week by default.

As for categories of news articles used for our system, we adopted eight categories: culture, society, government, economy, environment, education, sports, and science. These categories are frequently used in Web news sites of national daily papers in Japan such as Yomiuri, Nikkei, Asahi, Mainichi, and NHK.

3.4 Function of Hyperlinks of Locations

The information window has hyperlinks of all locations of the news article. Readers can

move to the related locations of the article on the map freely by using this function.

Figure 4 shows screenshots of moving the related locations by the hyperlinks.

Figure 4. Screenshots of moving the related locations by hyperlinks

4 EVALUATION

4.1 Outline of the Evaluation

As preliminary experiments of the proposed system, we conduct the experiment to evaluate the usefulness of interfaces and functions. The experimental subjects are 10 students from Kagawa University. We asked the subjects to use the proposed system about 10 minutes as a teacher for NIE in an elementary school. In order to compare the existing system and the proposed system, we asked the subjects to use Mapnews under the same situations, and to answer a questionnaire.

The questionnaire consists of three major items of "browsing news", "narrowing down news", and "comparison with Mapnews" based on 4-point Likert scale (4: Strong agree, 3: Agree, 2: Disagree, 1: Strong disagree), and the column of comments.

4.2 Evaluation of the Browsing News

The section 1 in Table 1 shows the evaluation results of "Browsing Function". All representative values except for the item 1-5 were higher than 3.0. As for the item 1-5, the value of average is 2.6 which is lower than that of other items. Therefore, we have to improve the design of the news list or supplement the information of the list when readers use this system.

4.3 Evaluation of Narrowing Down News

The section 2 in Table 1 shows the evaluation results of "Narrowing Down News". All representative values except for the item 2-4 were higher than 3.0. From the evaluation result of the item 2-4, we can see that the value of median is 2.6 and the value of mode is 2.0. Therefore, we can say that the subjects evaluated that the switch buttons for narrowing down by categories were not intuitive. We have to improve the design of the buttons.

4.4 Comparison with Mapnews

The section 3 in Table 1 shows the results of "Comparison with Map Newspaper". As you can see that all values were higher than 3.0. On NIE in elementary schools, we confirmed that our system is more useful than Mapnews.

4.5 Analysis of Comments

Figure 5 and **6** show the representative comments about the interface and functions of the proposed system. We can improve the interface based on the comments in Figure 5 easily.

As for the comments of functions in Figure 6, we can achieve all comment except for the last. In order to achieve the last comment, we have to consider how to collect the information about the principal products and the climate of the distinct area firstly.

ISBN: 978-1-941968-42-0 ©2017 SDIWC

Table 1. Experimental Results

1	Browsing News	avarage	median	mode
1-1	Items in the information window is appropriate.	3.4	3.5	4
1-2	Design of the information window is appropriate.	3.2	3	4, 3
1-3	Hyperlinks of all locations in the information window are useful.	3.3	3.5	4
1-4	Design of the news list under the map is appropriate.	3	3	3
1-5	News list is intuitive.	2.6	2.5	2
1-6	Display of the contents of the article is appropriate.	3.5	3.5	4, 3
1-7	Hyperlinks of all locations in the article is useful.	3.2	3	4, 3
2	Narrowing Down News	avarage	median	mode
2-1	Function of narrowing down news by keywords is appropriate.	3.3	3	3
2-2	Function of narrowing down news by a period of time is appropriate.	3.6	4	4
2-3	Function of narrowing down news by categories is appropriate.	3.2	3	4, 3
2-4	Design of the function of narrowing down news by categories is appropriate.	2.9	2.5	2
3	Comparison with Mapnews	avarage	median	mode
3-1	Categories of news are more useful for browsing than Mapnews.	3.9	4	4
3-2	Hyperlinks of all locations in the article is useful.	3.2	3	4, 3
3-3	Narrowing down news is easier than that of Mapnews.	3.8	4	4
3-4	Narrowing down by a period of time is more useful.	3.9	4	4
3-5	Narrowing down by categories is more useful.	3.7	4	4

- The supplementary explanation of hyperlinks of all locations in the news should be added.
- As for the interface of narrowing down news by categories, the checkbox is more intuitive than the switch button.
- It is not easy to confirm locations where the focus moved on the maps after clicking the hyperlinks of locations in the article.
- The date of issue of news should be displayed in the news list.

Figure 5. Representative comments of the interface

- Readers cannot set the maximum number of markers displayed on the maps.
- The function to cancel narrowing down news is necessary.
- The function of full text search is necessary.
- It would be better that the principal products and the climate of the display area are displayed on the maps.

Figure 6. Representative comments of the functions

5 Improvement of the proposed system

We improved the design of the interface of the proposed system based on the evaluation results. Figure 7 shows the improved interface of the system. As shown in Figure 6, the switch buttons to choose categories were improved to the checkbox based on the comments.

Figure 7. A screenshot of the improved main interface

6 CONCLUSION

In this paper, we proposed the prototype system to browse Web news using maps, in order to provide support for NIE in elementary schools in Japan. We confirmed the usefulness of our system through the preliminary evaluations. Then, we improved the interface of the proposed system based on the evaluation results.

In the future work, we are going to ask elementary school teachers to use our system and answer the questionnaire, in order to evaluate the usefulness of the proposed system. Then, we will improve the proposed system based on the evaluation results. After that, we will integrate the proposed system with our

ISBN: 978-1-941968-42-0 ©2017 SDIWC

other research for providing support for NIE, such as the method of text simplification and the method to search for image contents for supplementing the contents of news articles. Finally, we will make a comprehensive evaluation of the integrated system at elementary schools in Japan.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant Number JP16K00478.

REFERENCES

[1] NIE using Newspaper, http://nie.jp/ (in Japanese)

[2] K. Kobayashi and K. Ando, "A Keyword Extraction Method for Newspaper Reading Support System for Elementary School Students," IPSJ CE, Vol.2013-CE-119 , No.17 , pp.1-6 , 2013. (in Japanese)

[3] M. D. Lieberman and H.Samet, "Supporting rapid processing and interactive map-based exploration of streaming news", Proc. of the 20th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp.179-188, 2012.

[4] B. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, J. Sperling, "NewsStand: A new view on news", Proc. of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp.144-153, 2008.

[5] Mapnews, http://www.mapnews.jp/ (This site was closed. Confirmed on 30th June, 2017.)

[6] Google Maps API, https://developers.google.com/maps/.

[7] Japanese Named Entity Extraction API, https://labs.goo.ne.jp/api/jp/named-entity-extraction/

[8] A. Kuroda and K. Ando, "Investigation of Location Information for Web News Search Support based on Maps," SJCIEE2014 , p.329, 2014. (in Japanese)

Accelerated Moving Multi-Agent Behavior on Two Configurations

Yuta Tsuruoka
Graduate School of Science and Engineering,
Hosei University, Kajino-cho 3-7-2,
Koganei-shi, Tokyo 184-8584
Email: yuta.tsuruoka.8s@stu.hosei.ac.jp

Takahiro Suzuki
Faculty of Science and Engineering,
Hosei University, Kajino-cho 3-7-2,
Koganei-shi, Tokyo 184-8584
Email: takahiro.suzuki.5q@stu.hosei.ac.jp

Marin Numazaki
Faculty of Science and Engineering,
Hosei University, Kajino-cho 3-7-2,
Koganei-shi, Tokyo 184-8584
Email: marin.numazaki.7u@stu.hosei.ac.jp

Shihoko Tanabe
Faculty of Science and Engineering,
Hosei University, Kajino-cho 3-7-2,
Koganei-shi, Tokyo 184-8584
Email: shihoko.tanabe.8e@stu.hosei.ac.jp

Isamu Shioya
Faculty of Science and Engineering,
Hosei University, Kajino-cho 3-7-2,
Koganei-shi, Tokyo 184-8584
Email: shioyai@hosei.ac.jp

***Abstract*—This paper discusses the coordination of autonomous stochastic moving multi-agents on two kinds of resources consisting of cells; three agents on a line, and three agents on a circle. Each agent on their resources stochastically moves on the cell resources through moving paths with time-lag. Our problem is how to coordinate agents to maximize resource utilization efficiently, where each agent depends on other agents locating the neighbors. We consider a shake for agents in order to increase the resource utilization, and we also call the shake an acceleration. The most highest resource utilization is that every cells are always occupied by agents, i.e. it is desirable to be the fewest expected number of cells not occupied by agents. We show that the resource utilization of the multi-agents becomes higher if every agents have appropriate acceleration. The acceleration depends on the configurations of cell resources, and we can find an optimal acceleration which is to maximize the cell resource utilization depending on their configurations.**

Keywords— **Autonomous Moving Stochastic Multi-Agents, and Resource Utilization.**

I. Introduction

This paper discusses a resource utilization of autonomous stochastic moving multi-agents with time-lag, and their agents stochastically move on the cells through moving paths along finite resources. The resources arrange over a configuration according to transition probabilities in synchronization. We consider two kinds of configurations; three agents on a line, and three agents on a circle. The moving of each agent is restricted so that it depends on the number of other agents on current cells within specific ranged windows, and there is coordination among agents with time-lag depending the window sizes, while the coordination depends on the resource configurations. The interactions among the agents are complicated to analyze the behavior of them, since each agent stochastically moves over the resources autonomously. In addition, the acceleration makes our analysis even more difficult. The paper[22], [21] shows that the stochastic moving multi-agent behavior becomes more stable if every agents arranged on line resources take an acceleration as a shake. The appropriate acceleration makes the coordination of moving-multi-agents maximize. The paper [21] discusses the resource utilization of agents considering the boundary effect on line resources. In this paper, we consider two kinds of cell resource configurations, their agents are accelerated by a shake, and we present their theoretical analysis of cell resource utilization on two kinds of resources. Then, we assume the acceleration of every agents is independent from the locations of resources. Our analysis shows that the resource utilization to maximize them depend on their configurations. Stronger acceleration is necessary in proportion to the number of boundaries in order to increase the efficient use of resources, where the boundaries are cells moving path that is a stop, i.e. a dead end cell of paths.

In our real world, there are a lot of unusual beings with unexpected phenomena that are beyond human understanding. In fact, a multi-agent behavior is one of them, while it is quite difficult to analyze the behavior of multi-agents in general theoretical frameworks, because there are interactions or coordination among agents. Fortunately, letting you perform intensive analysis on your desk, when we discuss the small sizes of configurations.

The studies of complex systems[1] have been expected to explore new unexpected phenomena which are carried by natural or artificial systems. The most attractive one is that the behavior of entire systems does not obvious from a simple combination of each agent behavior. Our stochastic moving multi-agents or multi-objects that take an appropriate average moving speed exactly behave more stable moving or achieve high resource utilization.

We are in need of a simple model with no fat in moving multi-agents for analyzing complex systems. Fortunately, Sen et al.[19] proposed a simple model for analyzing the behavior of moving multi-agents, and Rustogi et al.[17] presented the fundamental results of the former model. Ishiduka et al.[8] also

ISBN: 978-1-941968-42-0

introduced a time lag and showed the relationship between time lag and stability in moving multi-agents. The above models are intended to clarify how fast the moving multi-agents fall into a complete stable state, i.e. a hole state in absorbing Markov chain[5], thus the goal is to design a coordinative system which falls into a stable hole in shorter passage time as soon as possible.

On the other hand, in physics, Toyabe et al. [23] experimentally demonstrated that information-to-energy conversion is possible in an autonomous single stochastic moving agent. In other words, the paper presented a solution of Maxwell's devil. The idea is that if an agent goes up the spiral stairs during stochastic movements, it sets the stopper on the stairs so that the agent does not to come down. This approach needs an explicit control that the agent does not come down the spiral stairs. It is a single agent against a multi-agent, and our ultimate goal is to get the energy from stochastic moving multi-agents. In multi-agent models, Hiyama[7] presented the precise theoretical calculation providing the interactions among different types of objects in nucleus. These are realized by stacking a small effect, and also this paper use the stacking of the diffusion in stochastic moving objects.

Our model, Multi-Agent behavior with Time lag and Moving Speed: MATMS, is based on Sen et al.[19] and the developed model with time-lag proposed by Rustogi and Singh[17]. We note that our purpose is different from the papers [19], [17], [8] which try to clarify the relationships between time-lag and stability in multi-agent systems. In other words, their papers try to find the multi-agent configurations satisfying autonomous uniform resource allocation in a shortest passage time. Our model satisfies Markov condition and irreducible so that the states do not depend on the initial states in the limit, and our problem is to find more stable multi-agent states accompanying agent movements. It just likes as a molecule has an energy so that it is always moving while the agents are alive, and it depends on the manner of substances. The paper[22] showed that a stochastic moving multi-agent system, whose the agents move slowly as a whole on average, is more stable than other ones not having average moving speed as a whole theoretically. This paper demonstrates higher resource configuration behavior when we consider the agent locations on resources, that is, boundary effects. We can find the basis of our model in Shilling model[18].

This paper is organized as the following. First, we define our model in Section II. Section III and IV show that there exists a behavior of multi-agents based on theoretical analysis on two configurations: a line and a circle. In the following section, we discuss the related works. Finally, we conclude this paper in Section VI.

II. Stochastic Moving Multi-agent model with time-lag and moving speed

We shall consider two kinds of finite cell resources in Figure 1, and three agents arranged on their cells. All the agents move over the resource consisting of cells $S(i)$, where the resource configurations are a line and a circle. The papers [19], [17] only consider on a circle. All agents run synchronously in discrete time over the resource according to the following transition probabilities $p_{i,j}$ in stochastic manner. In the following, sometimes we are simply expressed as i a resource $S(i)$.

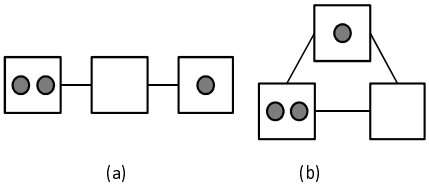


Fig. 1. Two kinds of resources and stochastic moving multi-agents.

First, we define a weight function $f_{i,j}$, $i, j = 1, .., n$ as

$$f_{i,j} = \begin{cases} 1, & i = j \\ 0, & i \neq j, r_i < r_j \\ 1 - \frac{1}{1+\gamma \exp(\frac{move(r_i - r_j, i, j) - \alpha}{\beta})}, & \text{otherwise,} \end{cases} \quad (1)$$

where r_i are the number of agents on i-th cell, and α, β and γ are constants. α is called an "inertia" which is the tendency of an agent to stay in its resource[17], and $move$ is an accelerated function to give average moving speed on either left or right defined by (2) and (3) in later. Rustogi et al. model[17] does not satisfy the condition "irreducible" exactly, while our model satisfies Markov property under the condition not to restrict the moving directions of agents, and the model becomes irreducible.

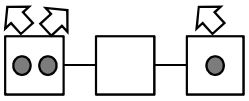


Fig. 2. The model MATMS.

Our model also has an average moving *speed* such as every agents move with an average moving speed s_i ($s_i \geq 1$) either of left or right directions on the cells, where i are i-th locations of agents. Their agents move along the resources arranged over the line or the circle according to the probability $p_{i,j}$ in stochastic manner, where i and j indicate i-th and j-th cells, respectively. In the case of right average moving speed s_i, the function $move(x, i, j)$, which describes the ratio of imbalance from a cell i to a destination cell j with the difference $x(= r_i - r_j)$ in the numbers of agents on i and j, is defined by

$$move(x, i, j) = \begin{cases} s_i \times x, & i < j, \\ x, & \text{otherwise,} \end{cases} \quad (2)$$

wheres s_i is an average moving speed at i-th cell. On the other hand, for the left average moving, $move$ is defined by

$$move(x, i, j) = \begin{cases} s_i \times x, & i > j, \\ x, & \text{otherwise.} \end{cases} \quad (3)$$

ISBN: 978-1-941968-42-0 ©2017 SDIWC

The function $move(x, i, j)$ is not symmetric with respect to x, and s_i represents the ratio of imbalance at i-th cell in the function $move$. As a special case, $move(x, i, j)$ becomes symmetric with respect to x and the agents do not move on average if $s_i = 1$. The average moving directions are inverted with the same average moving speed if each agent arrives at the leftmost or rightmost cells, i.e. the cells on the boundaries. The average moving speed depends on the agent locations, and each agent moves towards either left or right directions on average independently. Thus, all the agents are randomly choosing the moving directions which are apart from the effect on the left and right boundaries.

A moving transition probability $p_{i,j}$ from a current cell $S(i)$ to a destination cell $S(j)$ is defined by the normalization of $f_{i,j}$ with probability 1 as

$$p_{i,j} = \frac{f_{i,j}}{\sum_k f_{i,k}}, \quad i, j = 1, .., n, \tag{4}$$

based on $f_{i,j}$. Rustogi et al. [17] introduced a window $win(i)$ with a fixed size for analyzing the behavior of multi-agent systems with time-lag. Then, a moving transition probability $p_{i,j}$ from a current cell $S(i)$ to a destination cell $S(j)$ is defined by

$$p_{i,j} = \begin{cases} \dfrac{f_{i,j}}{\sum_{k \in win(i)} f_{i,k}}, & i = 1, .., n, j \in win(i), \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where w is a window size, and $win(i)$ is the set $[i-w, i+w]$. A time delay which is local properties is proportional to the window size w (see [17], [8]).

There are no constraints on the moving of agents such that each cell has a fixed upper limit capacity to occupy agents, while there is another constraint in the model, i.e. the moving transition probability $p_{i,j}$ is 0 if the number of agents on a cell $S(i)$ is less than the number of agents on a destination cell $S(j)$.

Our proposed model, Multi-Agent behavior with Time delay and Moving Speed: MATMS, is similar to the models [17], [8]. The resources in MATMS are arranged over a line $[1, 2, .., n]$ as in [8] or a circle[19], [17], and the wind function $win(i)$ is the set $[i-w, i+w] \cap [1, n]$ if w is a window size. We note that there are two choices on the moving average directions which are either left or right. Suppose an agent moves towards left on average at the previous step. Which is the moving direction at the next step? If we exclude the cases that the agents stay on boundaries, there are two exclusive cases(or a model protocol) for each agent independently: (1) we inherit the directions at the previous steps, i.e. left on average in above, or (2) we randomly select it at each step according to even probability either left or right, i.e. half to half rule for the direction. The second case (2) is suite to Markov property. The first case (1) does not satisfy Markov property so that the systems depend on the initial configurations.

III. Theoretical analysis of 3×3 model on the line

In this section, we present a concrete moving multi-agent such that the multi-agent taking an appropriate average speed achieves higher resource utilization than a multi-agent not taking moving average speed, i.e. every cells are occupied by agents in many cases on average.

Suppose the multi-agent of which the number of cells and agents are 3 together. This is a minimal model to examine a coordination among agents. We first use the parameter values $\beta = 2$ and $\gamma = 1$, and fix the window size w to 1.

Suppose 1, 2 and 3 are their cell names(Figure 1(a)) from left to right. We do not distinguish the names among the agents for the simplicity, and represent it just a. Suppose that the same average moving speed s_1, s_2 and s_3 are both s ($s \geq 1$), respectively, since the resource is symmetric. The moving directions of the agents are randomly selected either left l or right r in half and half at every steps. The multi-agent state is a set of three agent states. $[(a, 1), (a, 2), (a, 2)]$ is an example of the multi-agent states, where a are agents and 1, 2 and 3 are the resources.

An example of the agent moving configuration is represented by $(a, r, 1)$ if the agent a on the cell 1 moves towards right with average moving speed s_b. The multi-agent moving configuration consists of three agent configurations in this minimal model. For an example, $[(a, r, 1), (a, r, 1), (a, r, 1)]$ is a multi-agent moving configuration.

In our minimal model, the directions of average agent moving are stochastically chosen at every steps so that the multi-agent becomes Markov chain. In this setting, there are 10 multi-agent states shown in Figure 3, and we must consider 136 probabilistic transition rules. That is, the number of the states(Figure 3), the state transition rules (Appendix in [22]) and the multi-agent moving configurations (The top items of Appendix in [22]) are 10, 136 and 20, respectively. The illustration of the transition rule (d-3) is shown in Figure 4.

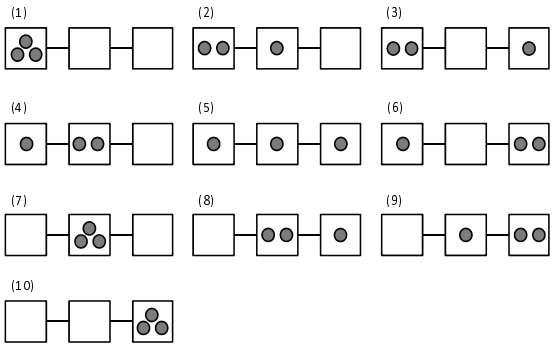


Fig. 3. The states of three agents arranged on the line consisting three cells.

This simple model satisfies Markov condition and it is irreducible, so we easily compute the eigenvectors of the state transition matrix with the size 10×10 using Appendix in [22], and compute the transition probabilities among every states

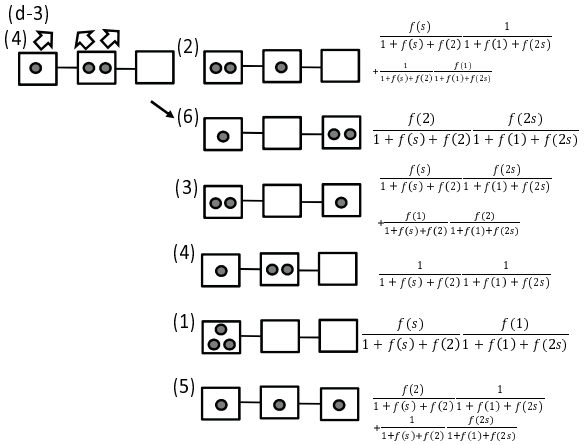


Fig. 4. The details of the transition rules (g-2) of the line, Appendix in [22].

TABLE I
THE PROBABILITIES STAYING THE STATES IN THE CASE $cells = 3$, $agents = 3$, $\alpha = 8$, AND $w = 1$ BASED ON THEORETICAL COMPUTATION OF THE LINE CONFIGURATION.

state	$s = 1$	$s = 7$
(1)	0.0003534811	0.001364175
(2)	0.0879075822	0.021569380
(3)	0.0906589070	0.027278335
(4)	0.0936260239	0.049030678
(5)	0.4540795018	0.793757492
(6)	0.0906589070	0.027278335
(7)	0.0008285100	0.007757373
(8)	0.0936260239	0.049030678
(9)	0.0879075822	0.021569380
(10)	0.0003534811	0.001364175

in the limit by changing the moving speed. The theoretical computational results of the existence probabilities for every states are shown in Table I.

The expected average number m_1 of the cell 1 occupied by agents is given by the following:

$$m_1 = p_1 + p_2 + p_3 + p_4 + p_5 + p_6, \quad (6)$$

where p_i are the probabilities of the correspondence states i shown in Table I. In other words, m_1 is the average resource utilization on the cell 1.

By similar way, we can compute the expected average numbers m_2 and m_3 of the cells 2 and 3, respectively, occupied by agents:

$$m_2 = p_2 + p_4 + p_5 + p_7 + p_8 + p_9, \quad (7)$$
$$m_3 = p_3 + p_5 + p_6 + p_8 + p_9 + p_{10}. \quad (8)$$

The whole expected average number v_m of cells occupied by agents, i.e. the resource utilization over the resource, is given as

$$v_m = (p_1 + p_7 + p_{10}) + 2(p_2 + p_3 + p_4 + p_6 + p_8 + p_9) + 3p_5.$$

TABLE II
THE EXPECTED AVERAGE NUMBER OF THE CELLS OCCUPIED BY AGENTS BASED ON THEORETICAL COMPUTATION OF THE LINE: $cells = 3$, $agents = 3$, $w = 1$.

speed	$s = 1$	$s = 7$
cell 1, mean m_1	0.8240108	0.9202784
cell 2, mean m_2	0.824081	0.942715
cell 3, mean m_3	0.8240108	0.9202784
resource, average vm	2.472103	2.783272

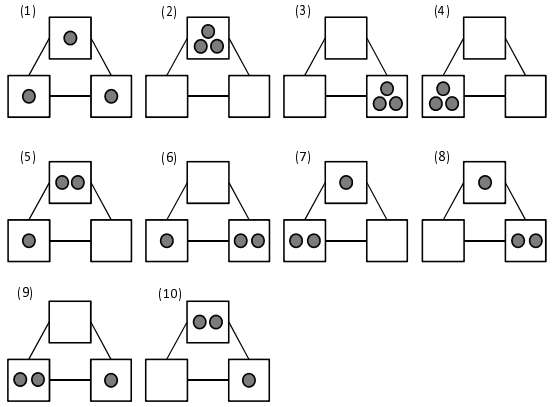


Fig. 5. The states of three agents arranged on the circle consisting three cells.

IV. THEORETICAL ANALYSIS OF 3×3 MODEL ON THE CIRCLE

In this section, we compute the multi-agent behavior without boundary effects on the circle: three cells and three agents shown in Figure 1, and we can compare with and without boundary effects. The number of states and the transition rules are 10 (shown in Fire 5) and 980 (one of them is shown in the Figure 6), respectively.

The resource utilization of the case can compute the ex-

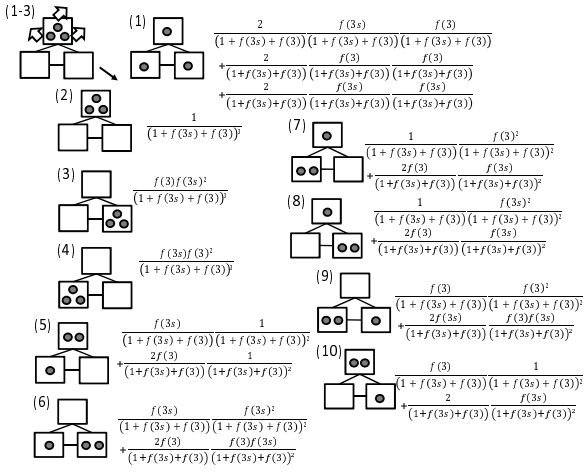


Fig. 6. The details of the transition rules (1-3) in the circle.

TABLE III
THE PROBABILITIES STAYING THE STATES IN THE CASE $cells = 3$, $agents = 3$, $\alpha = 8$, AND $w = 1$ BASED ON THEORETICAL COMPUTATION OF THE CIRCLE CONFIGURATION.

state	$s = 1$	$s = 7$
(1)	0.447229796	0.741141684
(2)	0.0007897928	0.002853046
(3)	0.0007897928	0.002853046
(4)	0.0007897923	0.002853046
(5)	0.0917333748	0.041716530
(6)	0.0917333748	0.041716530
(7)	0.0917333748	0.041716530
(8)	0.0917333748	0.041716530
(9)	0.0917333748	0.041716530
(10)	0.0917333748	0.041716530

TABLE IV
THE EXPECTED AVERAGE NUMBER OF THE CELLS OCCUPIED BY AGENTS ARRANGED ON THE CIRCLE, THEORETICAL COMPUTATION: $cells = 3$, $agents = 3$, $w = 1$.

speed	$s = 1$	$s = 7$
cell 1, mean m_1	0.8149537	0.9108608
cell 2, mean m_2	0.8149537	0.9108608
cell 3, mean m_3	0.8149537	0.9108608
resource, average vm	2.444861	2.732583

pected average numbers m_1, m_2 and m_3 of the cells 1, 2 and 3, respectively, occupied by agents:

$$m_1 = p_1 + p_2 + p_5 + p_7 + p_8 + p_{10}, \quad (9)$$

$$m_2 = p_1 + p_3 + p_6 + p_8 + p_9 + p_{10}, \quad (10)$$

$$m_3 = p_1 + p_4 + p_5 + p_6 + p_7 + p_9. \quad (11)$$

The whole expected average number v_m of cells occupied by agents, i.e. the resource utilization over the resource, is given as

$$v_m = (p_2 + p_3 + p_4) + 2(p_5 + p_6 + p_7 + p_8 + p_9 + p_{10}) + 3p_1.$$

Table V shows that there is an optimal speed to accelerate the MATMS in the circle 3×3 as well as the line configuration (also see [21] on a line case). But, the circle 3×3 is a boundary-less, so the resource utilization becomes low rather than boundary cases in the same given acceleration.

TABLE V
THE RESOURCE UTILIZATION IN THE CASE $cells = 3$, $agents = 3$, $\alpha = 8$, $\beta = 2$, $\gamma = 1$ AND $w = 1$ BASED ON THEORETICAL COMPUTATION OF THE LINE AND THE CIRCLE CONFIGURATIONS.

speed	line v_m	circle v_m
1	2.452544	2.444861
2	2.609133	2.561073
3	2.723913	2.665648
4	2.779706	2.723538
5	2.795896	2.742501
6	2.793445	2.741467
7	2.783272	2.732583
8	2.770548	2.722051
9	2.758217	2.712767
10	2.747893	2.705684

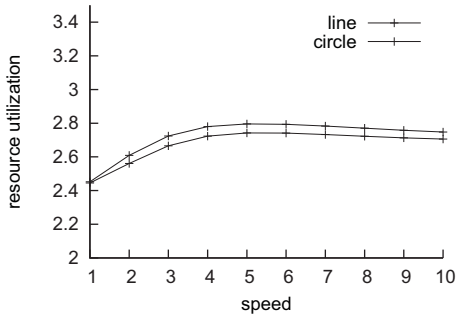


Fig. 7. The expected average number of cells occupied by agents based on theoretical computation in line and circle resource utilization: $cells = 3$, $agents = 3$ and $w = 1$, where $\alpha = 8$, $\beta = 2$, $\gamma = 1$.

V. RELATED WORKS

Sen et al.[19] presented our basic model, and Rustogi et al.[17] also proposed the their extended model with time delay and presented the excellent results. Ishiduka et al.[8] introduced a time lag for the propagation speed explicitly in addition to a window, and showed the relationships between stability and time lags. The similar result had been already obtained[9]. We note that Sen and Rustogi models employ the resources on circles. On the other hand, the resources of Ishiduka model are on a straight line. A straight line of resources are more realistic and natural compares to the circle. How's the boundary effect? How's the circular effect?

There are a lot of discussions on the stability of multi-agents. Chlie et al. [3] tries to find time Markov chains to be stable when its state has converged to an equilibrium distribution. Bracciali et al. [2] presents an abstract declarative semantics for multi-agent systems based on the idea of stable set. Moreau [14] discusses the compelling model of network of agents interacting via time-dependent communication links. Finke and Passino [4] discusses a behavior of a group of agents and their interactions in a shared environment. Lee et al. [10] considers the kinematicd based dynamics-based flocking model on graphs, and the model of the behavior is unstable. They proposed a stable information framework for multi-agents. Mohanarajah and Hayakawa [13] discusses the formation control of multi-agent dynamical systems in the case of limitation on the number of communication channels. Hirayama et al. [6] introduced the distributed Lagrangian protocol for finding the solutions of distributed systems. These papers are intended to control the multi-agent systems in corporative stable states. However, our model is one of the natural models to achieve the coordination without controls and without communication among agents.

VI. CONCLUSIONS

In this paper, we considered a stochastic moving multi-agent model, and presented that the model, Multi-Agent behavior with Time delay and Moving Speed: MATMS, having appropriate average stochastic moving speed become higher

resource utilization than ones not having average moving speed. Then, we considered the boundary effects on resources. This shows that each agent needs the moving acceleration to achieve hight resource utilization. The acceleration is a *shake* in this paper. In our model, we considered two kinds of resource configurations, the line and the circle, in moving multi-agents. Then, we showed how are boundary effects related to resource utilization in multi-agents.

REFERENCES

[1] R. Axelrod. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press, Princeton University, 1984.

[2] A. Bracciali, P. Mancarella, and K. Stathis. Stable multi-agent systems. *LNAI 3451*, pages 322–334, 2005.

[3] M. Chli, P.D. Wilde, J. Goossenaerts, V. Abramov, N. Szirbik, L. Correia, P. Mariano, and R. Ribeiro. Stability of multi-agent systems. *IEEE International Conference on Systems, Man and Cybernetics*, 1:551–556, 2003.

[4] J. Finke and K.M. Passino. Stable cooperative multiagent spatial distributions. *Decision and Control, European Control Conference*, pages 3566–3571, 2005.

[5] C.M. Grinstead and J. L. Snell. Introduction to probability. *American Mathematical Society*, 1997.

[6] K. Hirayama, T. Matsui, and M. Yokoo. Adaptive price update in distributed lagrangian relaxation protocol. *AAMAS*, 2009.

[7] E. Hiyama and et al. Three- and four-body cluster models of hypernuclei using the g-matrix an interaction. *Progress of Theoretical Physics*, 97(6):881–899, 1997.

[8] Y. Ishiduka and K. Iwanuma. A relationship between time delay and the amount of knowlege in distributed cooperation of multi-agent systems. *IEICE D-I*, J86-I:117–120, 2003. Japanese.

[9] Stuart Kauffman. *At Home in the Universe: The Search for Laws of Self-Organization and Complexity*. Oxford University Press, Oxford University, 1996.

[10] D. Lee and M.W. Spong. Stable flocking of multiple inertial agents on balanced graphs. *American Control Conference*, 2006.

[11] S. Lee and H. Wang. Multi-agent coordination using nearest-neighbor rules: Revisiting the vicsek model. *Cornell University*, 2004.

[12] V.R. Lessor. Reflections on the nature of multi-agent coordination and its implications for an agent architecture. *Autonomous Agents and Multi-Agent Systems*, 1:89–111, 1998.

[13] G. Mohanarajah and T. Hayakawa. Formation stability of multi-agent systems with limited information. *American Control Conference*, 2008.

[14] L. Moreau. Stability of multiagent systems with time-dependent communication links. *IEEE Transactions on Automatic Control*, 50, 2005.

[15] W. Ren, R.W. Beard, and E.M. Atkins. A survey of consensus problems in multi-agent coordination. *American Control Conference*, 2005.

[16] D. Robertson. Multi-agent coordination as distributed logic programming. *ICLP, LNCS 3132*, pages 416–430, 2004.

[17] S.K. Rustogi and M.P. Singh. Be patient and tolerate imprecision: How autonomous agents can coordinate effectively. *6th IJCAI*, 1999.

[18] Thomas C. Schelling. Dynamic models of segregation. *Journal of Mathematical Sociology*, 1:143–186, 1971.

[19] S. Sen, S. Roychowdhury, and N. Arona. Effects of local information on group behavior. *Proceedings of the Second International Conference on Multiagent Systems, AAAI*, 1996.

[20] O.M. Shehory, K. Sycara, and S. Jha. Multi-agent coordination through coalition formation. *4th International Workshop, ATAL'97*, pages 143–154, 1998.

[21] I. Shioya. An accelerated resource utilization in autonomous stochastic mobile multi-agents. *International Journal of Innovation in the Digital Economy*, 5:1–14, 2014.

[22] I. Shioya and T. Miura. An autonomous accelerated coordination of stochastic moving multi-agents under variations and resource utilization. *International Journal of Digital Information and Wireless Communications (IJDIWC)*, 2:943–957, 2012.

[23] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano. Experimental demonstration of information-to-energy conversion and validation of the generalized jarzynski equality. *Nature Physics*, 6:988–992, 2010.

ISBN: 978-1-941968-42-0 ©2017 SDIWC

Explicit but Stable Spring-Damper Model with Harmonic Oscillation

Young-Min Kang
Dept. of Game Engineering
Tongmyong University, Busan, KOREA
ymkang@tu.ac.kr

ABSTRACT

The mass-spring model is the most commonly used model for the animation of cloth and the physical simulation of various deformable objects. Although the method is simple, this model has, however, serious problems. The spring model tends to be numerically unstable. In other words, when you try to simulate a spring that does not stretch well, you will have to use very small time steps or to solve very large linear systems generated by implicit integration. In this paper, we propose a new simulation model that improves numerical stability and physical reality in order to perform large scale mass-spring model. The method can be easily parallelized. We implemented the method in CUDA environments which can be used on general personal computer. The proposed method is based on a harmonic oscillation model.

KEYWORDS

Mass-Spring, Stability, Harmonic Oscillation

1 INTRODUCTION

This paper proposes a model with improved numerical stability for efficient simulation of a mass-spring model, which is frequently used to perform physical simulations of deformable objects. The method of this paper reduces the numerical instability of the Euler method by applying method for better integration of spring force based on harmonic oscillator model. In general, the problem of numerical instability is resolved by applying implicit integration, but this method additionally causes a problem of solving a large-scale linear system. Especially, when the number of mass points increases, the number of components constituting the matrix of the linear system is increased to the square of the number of particle points. Therefore, the mass-spring model cannot be easily simulated in parallel fashion In this paper, a simulation method that can be easily implemented in CUDA environment was proposed. The CUDA environment is becoming more and more common in personal computer environments. Because of the parallelism and the numerical stability, the simulation method can efficiently produce plausible motions in real-time.

2 RELATED WORK

The mass spring model is a model that has been actively studied in the field of cloth animation since it was first formulated as a deformable object [11]. Among the various studies to solve the difficulties in real-time animation of deformable object model[12, 9, 10], the most important break-through was the implicit integration approach. Applying this implicit integration method to cloth animation, a linear system solution becomes a problem [1, 2]. The problem is that this linear system is a sparse matrix, but it has to deal with a very large matrix. A Several efficient schemes have been proposed to reduce the computational burden of the linear system solution including these matrices and use them for real-time cloth animation [8, 7, 4]. However, these techniques sacrifice accuracy to obtain real-time performance It was difficult for a complex model with a large number of particle points. A real-time animation technique of a character model fully dressed by Cordier has been proposed [3]. However, this technique did not improve the performance of the physics simulation it-

self. In this paper, we propose a method that can obtain stable results and calculate the state change of each material point in a form suitable for parallel processing.

3 PROBLEM

The simplest numerical integration technique is an explicit Euler integral. This method approximates the particle velocity of particles in the following way.

$$\frac{\mathbf{v}^{t+h} - \mathbf{v}^t}{h} \simeq \frac{d}{dt}\mathbf{v}^t = \frac{1}{m}\mathbf{f}^t \tag{1}$$

Therefore, the simulation is to calculate the velocity of the next state using the current velocity and the current force in the following manner.

$$\mathbf{v}^{t+h} = \mathbf{v}^t + \frac{h}{m}\mathbf{f}^t \tag{2}$$

This technique has serious numerical instability problems. This is because $\mathbf{f}^t h$ is not an accurate integration. To solve this problem, the implicit integration method performs the numerical integration using the force of the following state, that is, $\mathbf{f}^{t+h}$. However, the power of this future is unknown in the present state, and it is inevitable to use Taylor expansion. This approximation can be performed as follows.

$$\mathbf{f}^{t+h} = \mathbf{f}^t + \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\Delta\mathbf{x} \tag{3}$$

There is another problem that $\partial\mathbf{f}/\partial\mathbf{x}$ is not a vector. When we have n mass-points, $\partial\mathbf{f}/\partial\mathbf{x}$ becomes $n \times n$ matrix. Therefore, the problem is essentially a large linear system problem with $O(n^2)$ matrix.

Although implicit integration always guarantees a stable result, this large-scale linear system solution is difficult to implement and is not suitable for parallel processing. In order to solve this problem, we propose a new technique which has a greatly improved numerical stability and follows an explicit integration method which can be calculated independently for each particle. The basic idea is that the numerical instability was caused by an inaccurate numerical integral, which is to be corrected to perform a more accurate numerical integration. The accurate velocity change can be described as follows:

$$\mathbf{v}^{t_0+h} = \mathbf{v}^{t_0} + \frac{1}{m}\int_{t_0}^{t_0+h} \mathbf{f}^t dt \tag{4}$$

The problem is how to obtain this exact integral. Integrating the spring force defined in the Hooke's law results in a problem because it does not take into account the varying forces during the integration period. Therefore, we considered a model considering this. This model is based on the assumption that the movement of the spring will cause a harmonic vibration. Harmonic oscillation is described analytically and accurate integration is possible. Therefore, the problem is to obtain the value of equation 4 analytically.

To solve the problem, simple harmonic oscillation without the consideration of damping force was considered [6]. However, damping was not considered in the previous work. In this paper, we propose a new method that correctly integrates the spring forces taking the damping force into account, and the better time stepping strategy is also proposed.

The proposed method takes into consideration only the state of adjacent material points, as in the explicit integration method. This approach is well suited for parallelism. Therefore, we implemented the proposed technique in CUDA environment. When simulating particles in the GTX 590 GPU environment with

CUDA, it was possible to simulate a complex mesh of 16,384 particles with a performance of more than 66 frames per second.

4 HARMONIC OSCILLATION

Suppose two particles i and j are linked with a spring. The spring can be described as (i, j), and the locations of the particles are denoted as $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively. The velocities and masses are similarly described as $\mathbf{v}_i$, $\mathbf{v}_j$, m_i, and m_j.

The vector from i to j is $\mathbf{x}_{ij} = \mathbf{x}_j - \mathbf{x}_i$, and the relative velocity can be similarly described as $\mathbf{v}_j i = \mathbf{v}_j - \mathbf{v}_i$. Let us denote a normalized vector with hat as follows:

$$\hat{\mathbf{x}}_{ji} = \frac{\mathbf{x}_{ji}}{|\mathbf{x}_{ji}|} \tag{5}$$

The location of a particle is a function of time, and usually denoted as $\mathbf{x}_i^t$. Other vectors describing the physical states of particles are similiarty expressed. The rest length of a spring is denoted l^0, and the length at time t is l^t and equals to $|\mathbf{x}_{ji}^t|$. The difference between the current length and the rest length is denoted as $\delta^t = l^t - l^0$. The stiffness of the spring is κ.

A simple example of a harmonic oscillation is the situation where a mass point in 1D space is linked to a static object with a spring. The location of the mass point can be described with a scalar value x, and it oscillates in accordance with the equation $x = A \sin \omega t$ where $\omega = \sqrt{\kappa/m}$.

When two mass-points are linked, we can describe the elongation of the spring as follows:

$$\delta^t = A \sin \omega t = A \sin \left(\sqrt{\frac{\kappa(m_i + m_j)}{m_i m_j}} t \right) \tag{6}$$

In other words, the spring with two mass-points m_i and m_j can be described as a simple harmonic oscillation where a static object holds a dynamic mass-point of $(m_i m_j)/(m_i + m_j)$. If m_i approach to infinity (static), this mass expresion becomes m_j.

In order to determine A, energy conservation is applied. Total energy is the sum of kinetic energy and potential energy. The total energy of the spring is $\kappa A^2/2$. The potential energy is $\kappa \delta^{t2}/2$. The kinetic energy can be described as follows:

$$\frac{1}{2} \left(\frac{m_i m_j}{m_i + m_j} \right) \dot{\delta^t}^2 \tag{7}$$

Therefore, the energy conservation can be described as follows:

$$\kappa A^2 = \kappa \delta^{t2} + \left(\frac{m_i m_j}{m_i + m_j} \right) \dot{\delta^t}^2 \tag{8}$$

We can then find the amplitude A as follows:

$$A = \sqrt{\delta^{t2} + \frac{m_i m_j}{\kappa(m_i + m_j)} \dot{\delta^t}^2} \tag{9}$$

In order to solve the above equation, the derivative of the spring deformation should be computed ad follows:

$$\frac{d}{dt}\delta^t = \frac{d}{dt}(l^t - l^0) = \frac{d}{dt} l^t \tag{10}$$

Therfore, we can compute the derivative as follows:

$$\begin{aligned} \frac{d}{dt}\delta^t &= \frac{d}{dt}\sqrt{\mathbf{x}_{ji}^{t\,\mathrm{T}} \mathbf{x}_{ji}^t} \\ &= \frac{1}{2} \mathbf{x}_{ji}^{t\,\mathrm{T}} \mathbf{x}_{ji}^{t\,-0.5} \frac{d}{dt} \mathbf{x}_{ji}^{t\,\mathrm{T}} \mathbf{x}_{ji}^t \\ &= \frac{1}{2} \mathbf{x}_{ji}^{t\,\mathrm{T}} \mathbf{x}_{ji}^{t\,-0.5} (2 \mathbf{x}_{ji}^{t\,\mathrm{T}} \mathbf{v}_{ji}^t) \end{aligned} \tag{11}$$

ISBN: 978-1-941968-42-0 ©2017 SDIWC

$$= \frac{{\mathbf{x}_{ji}^t}^{\mathrm{T}}\mathbf{v}_{ji}^t}{\sqrt{{\mathbf{x}_{ji}^t}^{\mathrm{T}}\mathbf{x}_{ji}^t}}$$
$$= {\hat{\mathbf{x}}_{ji}^t}^{\mathrm{T}}\mathbf{v}_{ji}^t$$

The amplitude A can then be easily obatined.

The harmonic oscillation can be described with two time variables t and T as $\delta^t = A\sin\omega T$ where t is the normal time in the simulation world and T is the time defined in the oscillation period. Therefore, T ranges from 0 to $2\pi/\omega$. In order to correctly integrate the mass-spring model based on harmonic oscillation, T corresponding to the current time t must be computed as follows:

$$T = \frac{1}{\omega}\sin^{-1}(\delta^t/A) \tag{12}$$

Now, the spring force along the spring (i,j), f_{ij} causing the harmonic oscillation can be analytically integrated as follows:

$$\begin{aligned}\int_{t_0}^{t_0+h} {f_{ij}}^t dt &= \int_{t_0}^{t_0+h} \kappa\delta^t dt \\ &= \kappa A\int_{T_0}^{T_0+h}\sin\omega T dT \\ &= -\frac{\kappa A}{\omega}\cos\omega T|_{T_0}^{T_0+h}\end{aligned} \tag{13}$$

By exploiting the integration above, we can simulate the mass-spring model more accurately. Let us define a variable ϕ which is the velocity change during the time interval can be obtained by dividing the integrated force with total mass.

$$\begin{aligned}\phi_{ij} &= \frac{1}{2\omega(m_i+m_j)}\int_{t_0}^{t_0+h} f_{ij}^t dt \\ &= -\frac{\kappa A[\cos\omega(T_0+h) - cos\omega T_0]}{2\omega(m_i+m_j)}\end{aligned} \tag{14}$$

The velocity change ϕ_{ij} is distributed to the linked mass-points in accordance with their masses. Let us denote the velocity changes magnitude for the mass-points caused by the spring (i,j) as ν_i and ν_j. We can easily find the following relations:

$$\begin{aligned}\nu_i + \nu_j &= \phi_{ij} \\ \nu_i m_j &= \nu_j m_i\end{aligned} \tag{15}$$

We can finaly determine the velocity change of each mass-point as follows:

$$\begin{aligned}\nu_i &= \phi_{ij} mj/(m_i+m_j) \\ \nu_j &= \phi_{ij} mi/(m_i+m_j)\end{aligned} \tag{16}$$

The velocity change of each mass-spring is actually the sum of the all the velocity changes cause by linked springs. Therefore, the velocity change of each mass-point by spring oscillation, $d\mathbf{v}_i^s$, can be computed as follows

$$d\mathbf{v}_i^s = \sum_{(i,j)\in E}\frac{\phi_{ij} mj}{m_i+m_j}\hat{\mathbf{x}}_{ij} \tag{17}$$

where, E is the set of springs.

The velocity changes computed by integrating the oscillating spring forces makes the simulation more stable. The simulation steps can be described as follows:

5 DAMPING

Damping is usually employed for stability and more plausible animation. The integration scheme in the previous section did not take the damping into account. However, the damping can be easily incorporated. It is well-known that the damped spring oscillates as follows:

$$\delta^t = Ae^{-\alpha t}\sin\omega_d t \tag{18}$$

Algorithm 1: Simulation Steps

UpdateMass-SpringState
Data: dt: In
begin
  compute ϕ_{ij} for every spring (i, j)
  for *all particle* i **do**
    compute spring motion for particle i
      $d\mathbf{v}_i^s = \sum_{(i,j)\in E} \frac{\phi_{ij} mj}{m_i + m_j} \hat{\mathbf{x}}_{ij}$
    compute velocity change for particle i
      $\Delta\mathbf{v}_i = (d\mathbf{v}_i^s + \mathbf{f}_{external})h$
    update velocity for particle i
      $\mathbf{v}_i += \Delta\mathbf{v}_i$
    update location for particle i
      $\mathbf{x}_i += \Delta\mathbf{x}_i$

where the variable α is a coefficient propotional to the damping force, and ω_d is the dampened frequency of the oscillation because of the damping force. In the first subsection, phi will be computed, and α and ω_d will be computed in the following subsection.

5.1 Computing phi with damping

The force integration can then be rewritten with the damping as follows:

$$\kappa A \int_{t_0}^{t_0+h} e^{-\alpha t} \sin \omega_d t dt \tag{19}$$

With the assistance of calculus techniques, we can obtain the following form:

$$\kappa A \frac{e^{-\alpha t}}{\alpha^2 + \omega_d^2} (-\alpha \sin \omega_d t - \omega_d \cos \omega_d t)|_{t_0}^{t_0+h} \tag{20}$$

Therefore, the integrated force magnitude for spring (i, j) can be described as follows:

$$\begin{aligned}
&\kappa A \frac{e^{-\alpha t_0} e^{-\alpha h}}{\alpha^2 + \omega_d^2} \quad [-\alpha \sin(\omega_d t_0 + \omega_d h) \\
&\qquad\qquad -\omega_d \cos(\omega_d t_0 + \omega_d h)] \\
&-\kappa A \frac{e^{-\alpha t_0}}{\alpha^2 + \omega_d^2} \quad (-\alpha \sin \omega_d t_0 - \omega_d \cos \omega_d h) \\
&= \kappa A \frac{e^{-\alpha t_0} e^{-\alpha h}}{\alpha^2 + \omega_d^2} \quad (-\alpha \sin \omega_d t_0 \cos \omega_d h \\
&\qquad\qquad -\alpha \cos \omega_d t_0 \sin \omega_d h \\
&\qquad\qquad -\omega_d \cos \omega_d t_0 \cos \omega_d h \\
&\qquad\qquad +\omega_d \sin \omega_d t_0 \sin \omega_d h)) \\
&-\kappa A \frac{e^{-\alpha t_0}}{\alpha^2 + \omega_d^2} \quad (-\alpha \sin \omega_d t_0 - \omega_d \cos \omega_d h)
\end{aligned}$$

For the simplicity, let us denote as follows:

$$\begin{aligned}
C^t &= \frac{e^{-\alpha t_0}}{\alpha^2 + \omega_d^2} \\
C^h &= \frac{e^{-\alpha h}}{\alpha^2 + \omega_d^2} \\
s_\omega^t = \sin \omega_d t_0, &\quad s_\omega^h = \sin \omega_d h \\
c_\omega^t = \cos \omega_d t_0, &\quad c_\omega^h = \cos \omega_d h
\end{aligned} \tag{22}$$

The ϕ for a spring (i, j) can then be described as follows:

$$\begin{aligned}
\phi_{ij} = \quad & -\frac{\kappa A}{2(m_i + m_j)} C^t \\
& (\quad s_\omega^t [C^h \alpha c_\omega^h - C^h \omega_d s_\omega^h - \alpha] \\
& + \quad c_\omega^t [C^h \alpha s_\omega^h + C^h \omega_d c_\omega^h - \omega_d])
\end{aligned} \tag{23}$$

Now we can apply the previous algorithm to obtain mass-spring motion with damping based on the ϕ values.

5.2 Determination of α and ω_d

Simple harmonic oscillaction with damping force can be expressed as follows:

$$\ddot{\mathbf{x}} + \frac{k_d}{m} \dot{\mathbf{x}} + \frac{\kappa}{m} \mathbf{x} = 0 \tag{24}$$

where k_d is a damping coefficient.

Without damping, the frequency of the oscillation ω is $\sqrt{\kappa/m}$. In order to compute the dampened frequency, let us define a substitution value ξ as $d/2\sqrt{\kappa m}$. Then the equation of motion can be described as follows:

$$\ddot{\mathbf{x}} + 2\xi\omega\dot{\mathbf{x}} + \omega^2\mathbf{x} = 0 \tag{25}$$

The solution of the equation is known as follows:

$$\mathbf{x}^t = Ae^{-\xi\omega t}\cos(\omega\sqrt{1-\xi^2}t) \tag{26}$$

Here, $\xi\omega$ is α, and $\omega\sqrt{1-\xi^2}$ is the dampened frequence ω_d. As we have seen before, the frequency of oscillation by two linked mass-points is computed as $\omega = \sqrt{k(m_i+m_j)/m_i m_j}$. Therefore, we can determine the values as follows [5]:

$$\begin{aligned} \xi &= \frac{k_d}{2\sqrt{\frac{\kappa m_i m_j}{m_i+m_j}}} \qquad (27)\\ \alpha &= \xi\omega = \frac{k_d(m_i+m_j)}{2m_i m_j} \\ \omega_d &= \sqrt{\frac{\kappa(m_i+m_j)}{m_i m_j}}\sqrt{1-\frac{k_d^2(m_i+m_j)}{2\kappa m_1 m_2}} \end{aligned}$$

With these values, we can simulate the harmonic oscillation of the mass-spring-damper model.

6 EXPERIMENTS

The system used to implement this technique was an intel 3.47GHz i7 CPU using Microsoft's Window 7 operating system, 24G RAM, and GTX590 GPU environment. Although it is a 12-core system, it did not perform parallel processing on the CPU, and parallel processing was performed using the GPU.

In order to compare the stability of the harmonic oscillation-based integration with simple Euler integration of spring force, we used the experimental setting shown in Fig. 1. In this setting, we can control the masses of mass-points, and the stiffness of each spring.

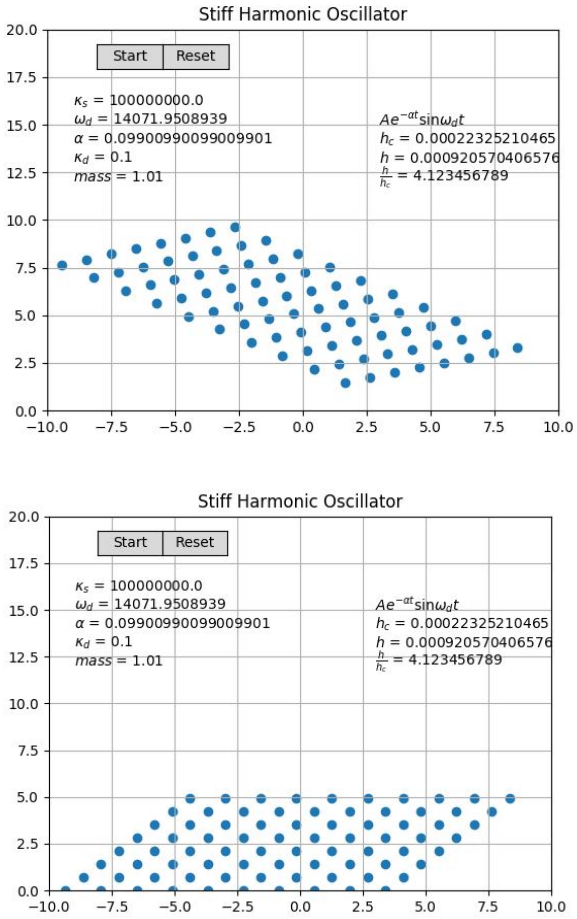


Figure 1. Experimental system

The experimental system integrated the spring force based on the harmonic oscillation model described in this paper, and showed more stable property compared traditional Euler method. In this simulation, we could increase the spring constant large enough to make the mass-spring model behaves almost like a rigid body as shown in Fig. 1.

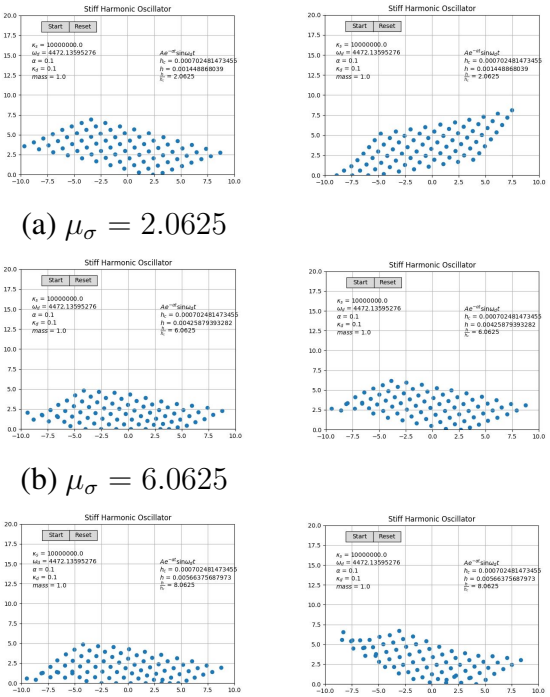
In order to understand the stability of the system, a critical time step should be defined. The motion of a spring is by its nature vibrating with the frequency of ω without damping or ω_d with damping. It is well known that we can represent the signal with a frequency ω by sampling at least twice in the cycling period, $2\pi/\omega$. Therefore, we have to reduce the time interval to be smaller than π/ω. We can define this time

step as a critical time step h_c

$$h_c = \frac{\pi}{\omega} \tag{28}$$

The traditional Euler method shows instability when the time step is larger than this critical time step. The proposed method can employ time steps larger than the critical time step. The stability of the system $\mu_\sigma$can then be measured with the ratio of the possible time step size to the critical time step as follows:

$$\mu_\sigma = \frac{h}{h_c} \tag{29}$$



(a) $\mu_\sigma = 2.0625$

(b) $\mu_\sigma = 6.0625$

(c) $\mu_\sigma = 8.0625$

Figure 2. Stability Test

Fig. 2 shows the stability of the proposed method. We employed large time steps to simulate the mass-spring model shown in the figure. Fig. 2 (a), (b), and (c) are the result when we changed the value μ_σ to be 2.0625, 6.0625, and 8.0625 respectively. As shown in the figure, even with the larger time steps compared with the critical time step, we could successfully simulated the model.

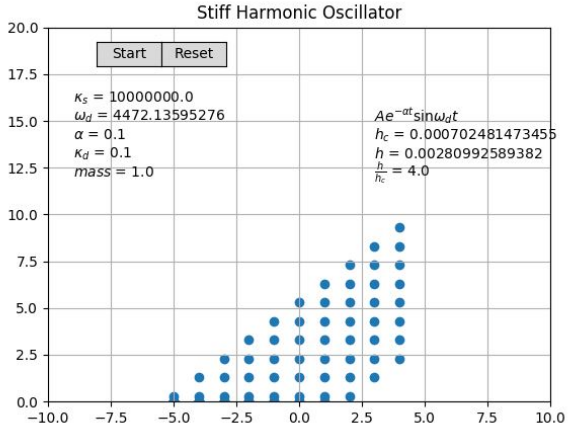


Figure 3. Integer mulitiplcation of h_c as time step

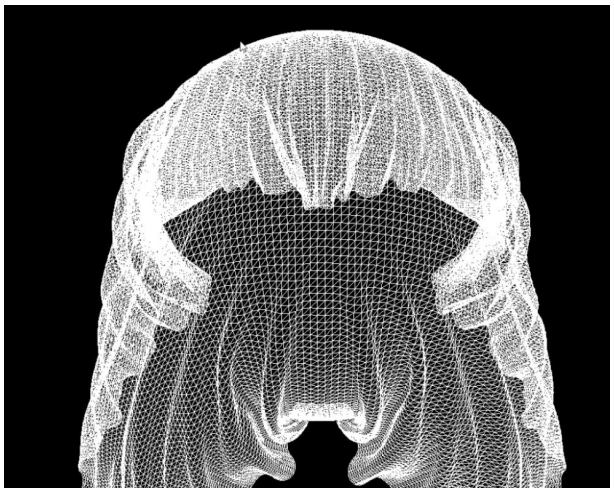
Fig. 3 shows the result when we use a time step which is an integer multiplication of critical time step. In this case, the integration of the force of spring becomes 0. Therefore, no spring motion is observed as shown in the figure. The proposed method can effectively compute the critical time step, and we can determine proper time steps based on the information.

The propose method is not only stable but also simple enough to be easily implemented in a parallel computing environments. We applied this method to a cloth animation and simulated the cloth in a parallel fashion with the assistance of GPU API. The result is shown in Fig. 4. As shown in the figure, the method can be successfully employed for massively large amount of mass-points. The number of mass-points used for this animation is 16,384.

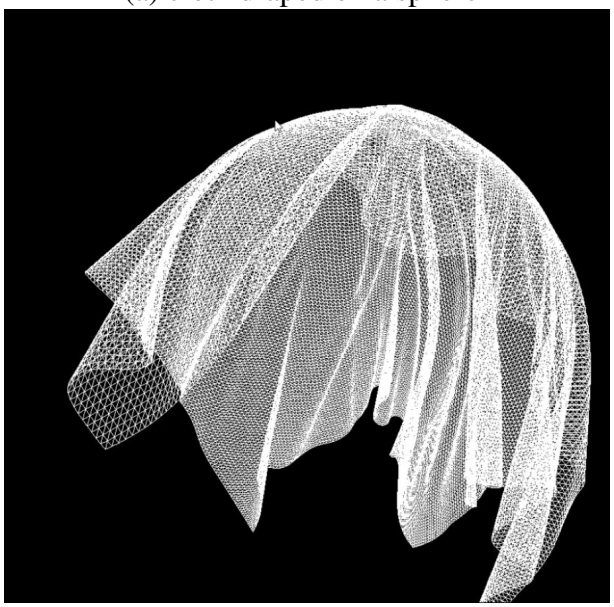
7 CONCLUSION

This paper proposed a simulation method of a mass-spring model with improved numerical stability based on a harmonic oscillation model. The proposed method is more stable than the simple explicit Euler method, and does not need to solve the linear system of the implicit method. These characteristics are suitable for parallel processing, and high-performance, high-quality simulation us-

ISBN: 978-1-941968-42-0 ©2017 SDIWC



(a) cloth draped on a sphere



(b) cloth sliding down on a sphere

Figure 4. Cloth animation with the proposed method

ing CUDA. The proposed technique can be implemented very simply in CUDA environment, and the experimental result is a technique to simulate the interactions of particles in massive mass-spring networks that were impossible in real-time applications.

ACKNOWLEDGMENT

This work was supported in part by 2017 PLSI Program of Korea Institute of Science and Technology Information.

REFERENCES

[1] D. Baraff and A. Witkin. Large steps in cloth simulation. *Proceedings of SIGGRAPH 98*, pages 43–54, July 1998.

[2] K.-J. Choi and H.-S. Ko. Stable but responsive cloth. *ACM Transactions on Graphics: Proceedings of SIGGRAPH 2002*, pages 604–611, 2002.

[3] F. Cordier and N. Magnenat-Thalmann. Real-time animation of dressed virtual humans. *Proceedings of Eurographics 2002*, 2002.

[4] M. Desbrun and M.-P. Gascuel. Animating soft substances with implicit surfaces. *Proceedings of SIGGRAPH 95*, pages 287–290, August 1995.

[5] D. H. House and J. C. Keyser. *Foundations of Physically Based Modeling and Animation*. CRC Press, 2016.

[6] Y.-M. Kang and C.-S. Cho. Photorealistic cloth in real-time applications. *Computer Animation and Virtual Worlds*, 23(3-4):253–265, 2012.

[7] Y.-M. Kang and H.-G. Cho. Real-time animation of complex virtual cloth with physical plausibility and numerical stability. *Presence - Teleoperators and Virtual Environments*, 13(6):668–680, 2004.

[8] Y.-M. Kang, J.-H. Choi, H.-G. Cho, and C.-J. Park. Fast and stable animation of cloth with an approximated implicit method. *Computer Graphics International 2000*, pages 247–255, 2000.

[9] M. Meyer, G. Debunne, M. Desbrun, and A. H. Bar. Interactive animation of cloth-like objects in virtual reality. *The Journal of Visulalization and Computer Animation*, 12:1–12, 2001.

[10] X. Provot. Deformation constraints in a mass-spring model to describe rigid cloth behavior. *Graphics Interface '95*, pages 147–154, May 1995.

[11] D. Terzopoulos, J. Platt, A. Barr, and K. Fleischer. Elastically deformable models. *Computer Graphics (Proceedings of SIGGRAPH 87)*, 21(4):205–214, July 1987.

[12] P. Volino, N. Magnenat-Thalmann, S. Jianhua, and D. Thalmann. An evolving system for simulating clothes on virtual actors. *IEEE Computer Graphics & Applications*, 16(5):42–51, September 1996.

Keyword diversity trend of consumer generated novels

Eisuke Ito
Research Institute for IT.
Yuya Honda
Grad. School of Library Science
Kyushu University
Motoka 744, Nishi-ku, Fukuoka, 819-0345, Japan.
{ito.eisuke.523@m, honda.yuya.128@s}.kyushu-u.ac.jp

ABSTRACT

Recent years, CGM (Consumer Generated Media), such as YouTube and nicovideo.jp for movies, syosetu.com for novel stories, become very popular. A lot of contents are posted to CGM sites every day, and also a large number of users are enjoying posted contents. At present, some articles mentioned decreasing diversity of contents. Some posted new content may be similar with previous posted contents. The authors are afraid that decreasing diversity of contents causes less energetic cultural activity. In this paper, the authors proposed two quantitative metrics of contents diversity, and applied them to the contents in syosetu.com. They focused the keywords which are given to the novel by the novel author, and calculated entropy and similarity of keywords. As the results, they observed increase of similarity, and it shows decrease of diversity of contents.

KEYWORDS

Big data analysis, CGM, word frequency, contents diversity, document term matrix, cos similarity.

1 INTRODUCTION

Recent years, CGM (Consumer Generated Media) services, such as YouTube and nicovideo.jp, are growing into social contents communication media. Not only movies but alto online novels are also popular. A lot of novels are posted to syosetu.com, and many users are reading and enjoying them every day. There are some major novel CGM sites such as syosetu.com, estar.jp, pixiv.jp and comico.jp. In this research, we focus on ”syosetu.com”, which is the most popular novel CGM site in Japan. There are more than 450 thousand online novels in syoteu.com as of Feb.2017. The number of contents and viewers are increasing. Most of online novels are written by amateur writers and might not be good quality. However, there are some high-quality novels and those popular high-quality novels are published as paper books. A few very popular novels may also become manga and animation.

We have been focused on syosetu.com as a research target of CGM contents search and recommendation. We proposed a ranking methods based on the analysis of novel keywords in [7], and reported the structure analysis of bipartite graph between user and contents in [2].

Nowadays, some blog articles mentioned that the diversity of novels may decreases in syotetu.com. Mr. Kawakami, who is the president of Dwango Company, mentioned that page view popularity ranking might cause decreasing diversity of CGM contents [3]. We believe that contents diversity is necessary to keep CGM site activity, and for cultural sustainability.

In this paper, we propose two quantitative metrics of contents diversity. One is entropy based metrics, and the other one is similarity based metrics. Especially, we focused on words in keyword field. Keywords are given to the novel by the novel author. We applied our metrics to keywords of novels, and evaluate our proposed metrics.

The rest of this paper is organized as follows. Section 2 shows basic statistics of novels and words on syosetu.com briefly. In section 3, we introduce an information entropy based quantitative metrics to measure contents diversity. Section 4 describes cosine similarity based contents diversity. The time series trend of the metrics indicates decreasing of contents diversity quantitatively. Finally, we conclude this paper in section 5.

ISBN: 978-1-941968-42-0

2 STATISTICS OF SYOSETU.COM

This section shows some statistics of novels and words on syosetu.com.

In this section, we explain the organization of the site syosetu.com, the number of novels and the number writers. We also describe a basic statistics concerning to the distribution of frequencies of words which were assigned to novels by users.

2.1 Novel metadata

Figure 1 illustrates a model of data structure and data flow of a novel in syosetu.com. When an author posts a novel to syosetu.com, the author can set the title, the author's name, genre, keywords and outline text as metadata. Author has to select a genre word from 15 genre words provided by the site. The author can give arbitrary words for keywords. Each novel is automatically assigned a novel ID, and novel ID is called as "Ncode" in syosetu.com.

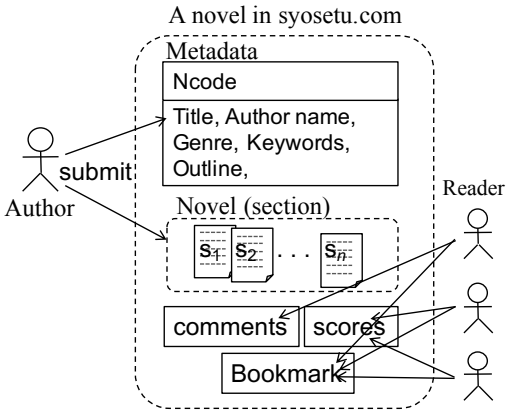


Figure 1. Metadata of syosetu.com

A novel in syosetu.com is classified into short novel or series novel. A serial novel consists of several sections, a short novel has one section. A novel of syosetu.com is also classified into completed novel or not. Short novel is certainly completed novel. Figure 2 shows the metadata page and the TOC (table of contents) page of a novel "Knight's & Magic" (Ncode: `n3556o`). This is a serial novel, and the TOC page shows section titles.

2.2 Novel metadata crawling

Web API named "Narou API" [4] is provided by syosetu.com. Using Web API, anyone can get novel metadata and author's bibliography (list of novels). Returned data is written in YAML format, and YAML is a format of structured data.

If you specify an Ncode in API, then you will get YAML format metadata of the novel. But we didn't use Ncode for metadata crawling, because Ncode assignment rule is not clear. On the other hand, author ID is a numerical number and assignment rule is very clear, that is serial number. Author ID (User ID) is assigned to each user in syosetu.com.

We created author's bibliography data crawler by Ruby language. The crawler increases the author ID number from 1 to 450,000, and gathers bibliography of the author ID. Since we checked in advance that the highest author ID may be about 400,000, we set 450,000 as upper limit of author ID to the crawler.

In order to efficiently crawling of 450,000 bibliography, we set 8 virtual machines for distributed parallel metadata crawling. We ran 8 crawlers since Nov. 2015 until Dec.2015. After crawling, we get 81,449 valid data. In other words, there are 81,449 authors in syosetu.com at Dec.2015.

Next, we created an extractor program. It extract each novel metadata (identified by Ncode) from 81,449 crawled author's bibliography data. Finally, we obtained 232,096 novel metadata. Table 1 shows the summary of data obtained by crawling and extraction. Table 2 shows a part of attributes of novel metadata.

Table 1. Crawled Metadata

Item	Description
Priod	Apr.2004–Oct.2015.
Novel	232,096
Writer	81,449

2.3 Novel posting trend

Figure 3 shows the number of monthly new novel posts. Until March 2014, the number of

Table 2. Attributes of metadata

Attribute	Description
ncode	Identifier of a novel
title	Novel title
story	Story outline of the novel
writer	Author name
keyword	Keyword(s) given by the author
genre	Genre word
writer	Author ID (User ID)
general_fistup	The first upload date

new novel posting is increase. Peak is March 2014. 5,306 new novel posted at March 2014. After that, new novel posting decreased, but more than 2,000 novels are newly posted to syosetu.com in a month.

2.4 Keyword trend

We counted the number of words in keyword field. Figure 4 shows monthly trend of the number of keywords. Blue line is total word count, and red line shows the number of unique words. As same as the number of new novel posting, the number of words is increase until March 2014. Peak is also March 2014. There are 7,149 unique words at March 2014. After that, the number of words decreased.

2.5 Keyword rank-frequency

We counted term frequency of each keyword. Figure 7 shows rank frequency plot of keywords. Both axes are in log scale.
Figure 7 illustrates a straight line in both log scale, therefore, the distribution of tag frequency follows the power-law distribution. We know that the word frequency in natural language documents follows the power-law distribution. Then, tags distribution is similar to natural language distribution.

3 ENTROPY BASED DIVERSITY

As we mentioned in section 1, some articles [4] mentioned diversity decrease of novels in syosetu.com. To investigate whether the diversity decreases or not, quantitative metric is necessary.
We use the following symbols for expressions.

D : a document (content) set,
n : the number of documents ($|D| = n$),
W : word set,
$df(w)$: document frequency of word w ($w \in W$).

3.1 Basic idea of entropy based contents diversity

Before definition of quantitative metrics, let us consider two extreme cases. Let n be the number of contents (the number of documents), and $df(w)$ be the documentary frequency of word w. If documents are perfectly uniform, all document are same. In this perfect uniform case, the same words will be given to all documents, and then, $df(w)$ will be n for all w.
Next, let us consider the opposite extreme case. If all contents were perfectly diverse, there is no similarity between any two contents. In this perfect diverse case, a word will be given to only one document, and then, $df(w)$ will be 1 for all w.
Actuary, $df(w)$ of a word w is between two extreme cases. Keywords, which are used as genre or category, are frequently appeared, and df of those words are high. Words, which represent creator nickname or content name, are appeared a few times, then df of those words are low.

3.2 Definition of tag entropy

Shannon estimated the entropy of real English documents[6]. He applied the information entropy to the words of document. The information entropy is calculated by the following expression (1). The unit of $H(W)$ is bit/word, if the bottom of a logarithm is 2.

$$H(W) = -\sum_{w \in W} p(w) log(p(w)) \qquad (1)$$

In expression (1), $p(w)$ is the appearance probability of word w. In syotetu.com, one word can be given at most one time for one novel. Therefore, $p(w)$ of a keyword w is $df(w)/n$.

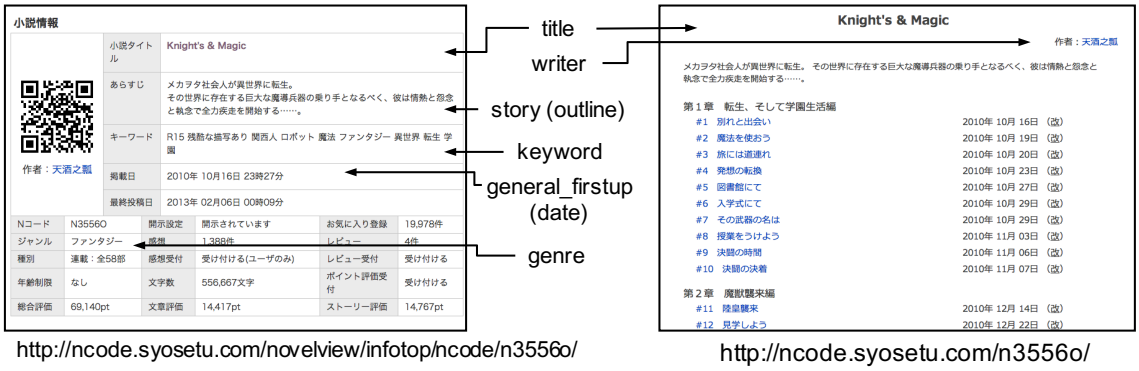


Figure 2. Novel TOC and metadata page (ID:n3556o)

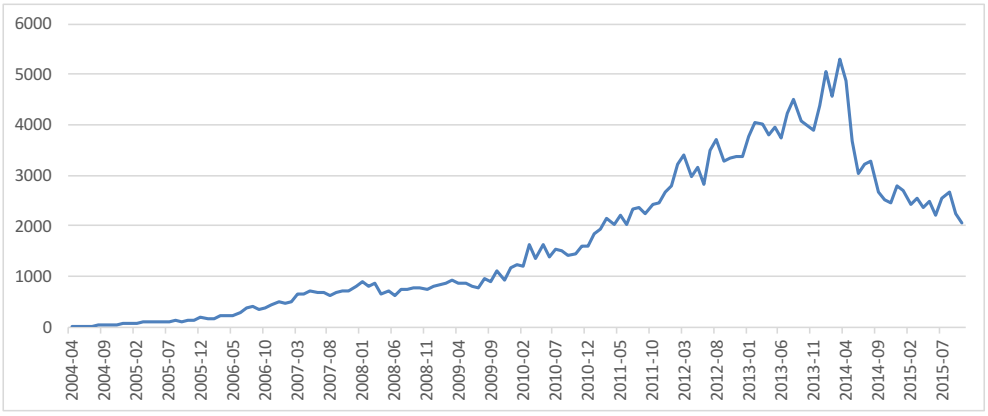


Figure 3. Number of posted novel (monthly)

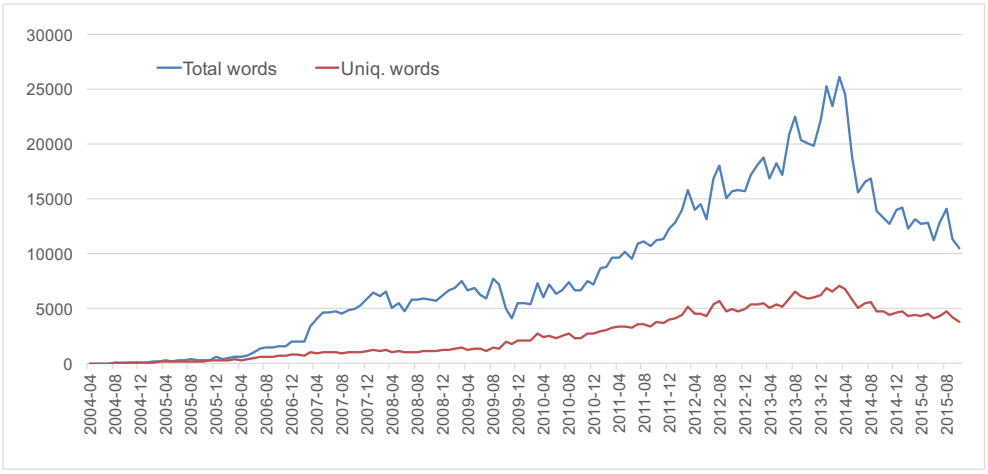


Figure 4. Number of (unique) words in keyword field (monthly)

3.3 Entropy trend of the keywords

Let D_m be the novel metadata document set, and all novel in D_m is posted at a specific month m of a year. For example, $D_{2010-01}$ is the novel metadata set, and all novel $d \in D_{2010-01}$ is posted at January 2010. We made the keyword set W_m extracted from D_m, and calculated $H(W_m)$ of keywords according to expression (1). Figure 5 shows the monthly trend of keyword entropy.

In Figure5, except during April 2017 to September 2009, the entropy of keywords gradually increases, and saturates about 10 bit/word. As shown in Figure 3 and Figure 4, the number of new novel posting and the number of (unique) keywords is increase until March 2014. After that, new novel posting and the number of (unique) keywords gradually decreased. Figure5 shows that the number of documents are not related entropy of word.

This result is similar with Shannon's result[6]. In [6], entropy of word is saturated about 11.8 bit/word for English documents.

During April 2017 to September 2009, entropy of keywords dented. This phenomenon will be mentioned in 4.4.

4 SIMILARITY BASED DIVERSITY

Secondly, we propose a content diversity metrics using cosine similarity.

4.1 Basic idea of entropy based contents diversity

Figure 6 illustrates a model of contents diversity. One dot in Figure 6 corresponds to a document. If documents are diverse, then distance between two documents will be long and similarity of the two documents will be small. On the other hand, if contents are not diverse, then distance of two contents will be close, and similarity of them are large.

There are some definitions for distances and similarity, such as Euclidean distance, Manhattan distance, and so on for distance, cosine similarity, Pearson's correlation, Jaccard/Dice/Simpson coefficient for similarity. In this study, we decided to use cosine similarity because it is most used as index of similarity. So, sum of all similarity of all document pairs may be a diversity metrics for document set.

4.2 Definition of $SumCos$

We use term-document matrix to vectorize a document. In the term-document matrix, a document is expressed as a word vector. The model is also known as "Bag of Words" model. Figure 7 illustrates an example of term-document matrix M. Usually, $M_{i,k}$ element is term frequency of word w_k in document d_i.

Cosine similarity of document i and j is calculated by expression (2).

$$cos(i,j) = \frac{\sum_k M_{i,k} * M_{j,k}}{\sqrt{\sum_k M_{i,k}^2} * \sqrt{\sum_k M_{j,k}^2}} \qquad (2)$$

The range of cosine is from -1 to 1, normally. In case of term-document matrix, the range of cosine similarity for every two documents or every two words, is 0 to 1. Because, all elements in the matrix are non-negative integer.

In case of syosetu.com, one keyword can be given at most one time for one novel. Then, an element $M_{i,k}$ must be 0 or 1 for keywords. Consequently, it is easy to calculate cosine similarity.

We calculated sum of cosine similarity ($SumCos$, for short) of all document pairs, according to the expression (3).

$$SumCos(D) = \sum_{i=1}^{n-1} \sum_{j=i}^{n} cos(i,j) \qquad (3)$$

For document set D, the number of pairs is ${}_nC_2 = n(n-1)/2$, where n is the number of documents ($n = |D|$). As shown in Figure 3, the number of posted documents (novels) in a month is different. If n is large, then the number of document pairs is more large, and $SumCos$ become larger value.

To modify the effect of the number of document, we propose $SumCos_{ave}$ in expression (4).

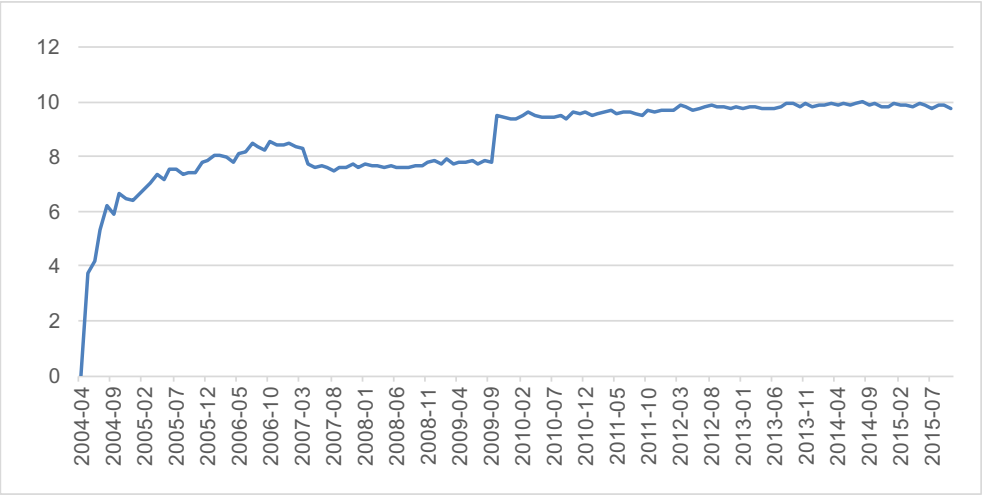


Figure 5. Entropy of keywords (monthly)

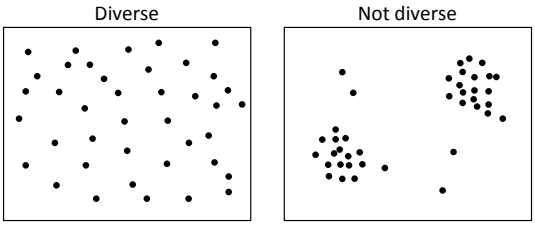


Figure 6. A model of diversity

word (term)

document	w_1	w_2	...	w_k	...	w_m
d_1	2	5		0		1
d_1	0	1		2		5
:				:		
d_i	1	1	...	M_{ik}	...	2
:				:		
d_n	1	0		0		10

Figure 7. Term-Document Matrix

$$SumCos_{ave}(D) = \frac{SumCos(D)}{{}_nC_2} \quad (4)$$

4.3 Monthly trend of $SumCos$

Let D_m be the novel metadata document set of month m of a year. According to the expression (4), we calculate $SumCos_{ave}(D_m)$ for each month. Figure 8 shows monthly trend of $SumCos_ave$ of keywords.

Except during April 2017 to September 2009, Figure8 shows that $SumCos_{ave}$ gradually increases. It indicates that the number of words commonly appearing in the keyword field of the novel has increased. This also indicates that the diversity of the novel keywords is decreasing.

4.4 Exception

During April 2017 to September 2009, entropy of keywords dented in Figure 5, and $SumCos$ of keywords is bumped in Figure 8. This period is exception of keyword data. We calculated ave_m with $ave_m = |W_m|/|D_m|$, where ave_m is average of the number of keywords per novel in month m. Figure 9 shows the monthly trend of ave_m.

Figure 9 obviously shows that ave_m is high during April 2017 to September 2009. Actually, keywords are increased in this exception period.

5 RELATED WORK

There are very few contents diversity trend analysis. Igami and Saka proposed a scientific paper mapping method called ”science map” [5] and investigate effectiveness of scientific policy of Japan. Science map is based on citation relations. A few top cited papers are core papers. The paper citing a core paper is a descendant paper. The descendant papers surround the core paper, and forms an island centered on the core. They reported evidence

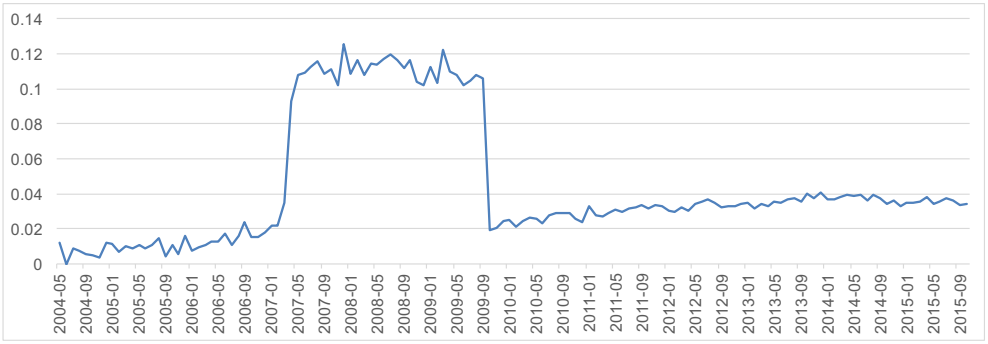


Figure 8. Average $SumCos_{ave}$ of keyword (monthly)

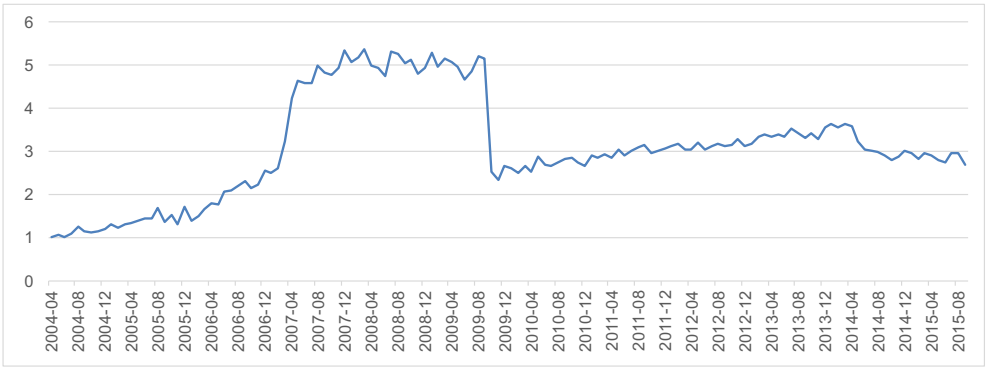


Figure 9. Average of the number of keywords (monthly)

of decreasing diversity of scientific article created by Japanese researchers[1]. Science map method is difficult to apply CGM contents. Because there are only few obvious direct links between novels.

6 CONCLUSION

Online novel become very popular in recent years. Some people mentioned decreasing of diversity of CGM contents. In this paper, we proposed two quantitative metrics for contents diversity. One is entropy-based diversity, and the other one is the sum of cosine similarity ($SumCos$). We applied proposed metrics to the set of syosetu.com novel metadata. The entropy of keywords is saturated about 10 bit/word. Word entropy is not suitable contents diversity metric.

The average of sum of the cosine similarity ($SumCos_{ave}$) is increased. It indicates increase the number of words commonly appearing in the keyword field of the novel. These may be quantitative evidences of decreasing contents diversity in syosetu.com.

In the future, we want to investigate trend of $SumCos$ not only for the keywords of the novel but also for the outline and the title. We want to investigate $SumCos$ trend of subsets for each genre, whether $SumCos$ trends are different or same. We also want to check diversity trend of other CGM contents such as movies or comic (cartoon). Finally, we want to establish user's contents selection model.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 15K00451.

REFERENCES

[1] Igami, M., Saka, A.: Decreasing diversity in japanese science, evidence from in-depth analyses of science maps. Scientometrics 106(1), 383–403 (January 2016)

[2] Ito, E., Shimizu, K.: Frequency and link analysis of online novels toward social contents ranking. In: Proc. of SCA2012 (The 2nd International Conference on Social Computing and its Applications). pp. 531–536. IEEE (November 2012)

[3] Kawakami, N.: Nico Nico Philosophy. Nikkei BP (November 2014)

[4] Narou-Developer: Narou api. http://dev.syosetu.com/man/api/

[5] Saka, A., Igami, M.: Science map 2010&2012 (study on hot research area (2005 – 2010 and 2007-2012) by bibliometric method). Tech. Rep. 159, NISTAP (July 2014)

[6] Shannon, C.E.: Prediction and entropy of printed english. Bell System Technical Journal, 30(1), 50–64 (1951)

[7] Shimizu, K., Ito, E., Hirokawa, S.: Predicting future ranking of online novels based on collective intelligence. In: Proc. of ICDIPC2013 (The Third Int'l Conf on Digital Info. Processing and Communications). pp. 263–274. IEEE (January 2013)

ISBN: 978-1-941968-42-0

An Attempt for Visual Design based on some kinds of data from Dynamic Statistics in Shikoku District

Yusuke Nakano Yoshio Moritoh* Yoshiro Imai**
General Incorporated Association Machitere
1-16-13 Saiho-cho Takamatsu Kagawa pref. 760-0004, Japan
*Department of Management Information, Kagawa Junior College
*1-10 Utazu-cho, Ayautagun Kagawa pref. 769-0201, Japan
**Graduate School fo Engineering, Kagawa University
**2217-20 Hayashi-cho Takamatsu Kagawa pref. 761-0396, Japan
E-mail: y.nakano@dex.co.jp, *moritoh@kjc.ac.jp, **imai@eng.kagawa-u.ac.jp

ABSTRACT

Recently it becomes more popular and more useful to access open data from government and to utilize such information for our daily lives. This paper describes how to visualize such open data from government and how to demonstrate importance of techniques for data processing and graphical manipulation. Aim of this study is to build 3-D modeling objects, to manipulate such objects on the major browsers not only of PCs but also of Tablets and to provide feasible environments for visualization of open accessible data through Web-based application.

KEYWORDS

Visual design, Open data, Web-based application, 3-D modeling objects.

1 INTRODUCTION

Recently a mount of open accessible data becomes more and more so that there will be a trend for some kind of people to handle and utilize such data for their lives, for example, learning, decision-making, management and future business if they could. In general, however, such open data may be only bulk of information for normal people so that it is necessary for them to recognize whether such data are important or not, namely almost all the people need the suitable methodology how to handle those open data effectively and efficiently. We think that visualization is one of the key solutions for above themes.

Visualization seems to be one of the useful and important technologies which can support even beginners of the relevant domain to understand outline of the related situation/phenomenon/problem at glance. For example, there have been some researches about visualization, in such cases it is reported that the relevant users have understood more than before with suitably visualizing schemes[1][2]. Our study is to utilize this visualization technology for effective and efficient handling of open accessible data from official body/public institution/local governments. In this paper, we describes how to visualize such open data from government and how to demonstrate importance of techniques for data processing and graphical manipulation.

This paper introduces related work for our more understanding in the next section. It illustrates our attempt of visual design based on some kinds of open accessible data from dynamic statistics in Shikoku district and its implementation in the third section. It explains our future application with our visual design in the fourth section. And finally it concludes our summaries in the last section.

2 RELATED WORK

For example, people sometimes refer the URL of "**re: form**"[3]. This URL says, "Rather than display the housing sales price data as a heat map, we chose to create a detailed map of the citys 325 neighborhoods and to extrude them based on the sales price per square foot - sort of a bar chart overlaid upon a city map. We felt that, visually, seeing the difference in heights

on a 3D map(shown in Figure1) would make the data more accessible and compelling."

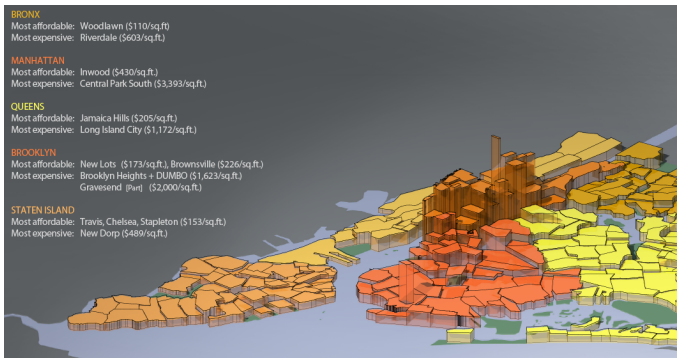


Figure 1. Seeing the difference in heights on a 3D map.

And it additionally mentions, "This map(shown in Figure1) helped us recognize how diverse the citys housing markets really were. They ranged from $110 per square foot in Woodlawn, Bronx to over $3,393 per square foot around Central Park South. Manhattan itself ranged from $430/sq.ft in Inwood to over $3,000 in NoLiTa. The 3D map shows the decline in prices as commute times increase. As one moves up the island: $1,600/sq.ft for the Upper West Side, to $800/sq.ft. for Morningside Heights, to $617 in Hamilton Heights, to $430 in Inwood. The 3D visualization also highlights several surprisingly affordable neighborhoods very close to Manhattan. Parts of Jersey City are about a 10 minute commute to the Village, and are about $450/sq.ft. And for comparison, the comparably-priced neighborhoods in Brooklyn or Queens would have a 3040 minute commute." The relevant map exactly **visualize** cost of the area at glance. Thank to the above map, people, who want to live at and/or borrow/lent suitable locations in New York city, will be able to determine where they obtain with their reasonable cost very easily and through an objective perspective.

3 IMPLEMENTATION OF SYSTEM

This section describes what we can obtain as open accessible data from local government and how to implement our visual design for such open data. It illustrates what we can obtain as a newly improved viewpoint through our attempt of visual design based on some kinds of open accessible data from dynamic statistics in Shikoku district.

3.1 Open Accessible Data

We have obtained the following graph for population for major municipalities in Shikoku district shown in Figure2. And we have also

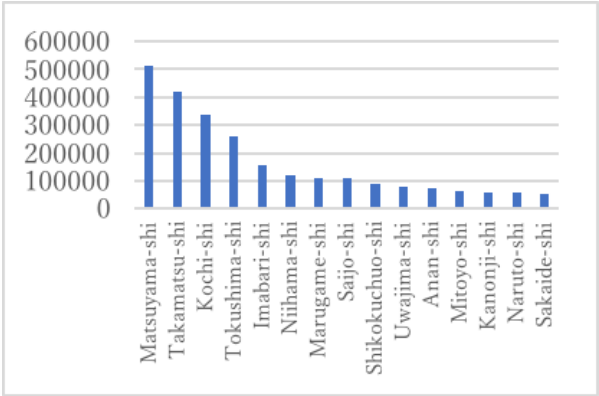


Figure 2. Top 20th population of municipalities in Shikoku (in 2015).

obtained the following graph for their area for the relevant municipalities in Shikoku district shown in Figure3.

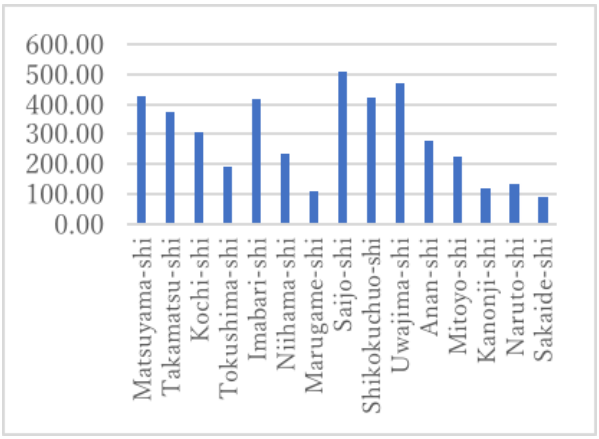


Figure 3. The relevant AREA of the same municipalities in Shikoku (in 2015).

The above data are to be manipulated by users but normally those information about population and area will be separately manipulated and processed for different aims.

ISBN: 978-1-941968-42-0 ©2017 SDIWC

3.2 Our Aim and System Implementation

As the related work above, we believe that the map together with other add-on information will become more useful and applicable not only for existing aim than previously but also for possible utilization even near future. So we have measured the map information and generated map-based data for future application from existing map information shown in Figure4.

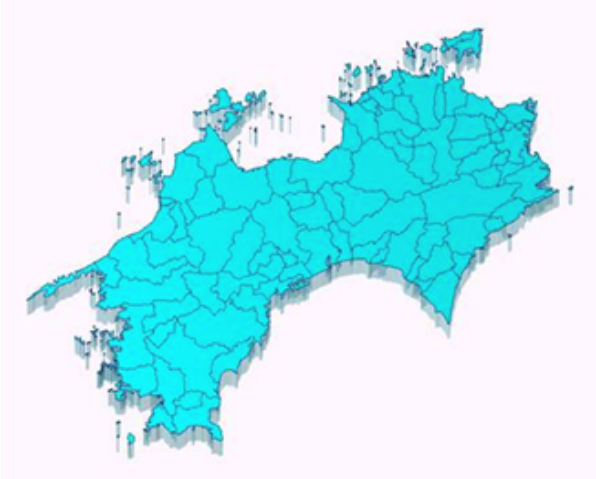


Figure 4. Displaying OBJ file on WebGLfrom an existing map of Shikoku District.

At the beginning, we have performed 3D modeling for all the municipalities in Shikoku. Our specification to design is as follows;

- Utilizing WebGL without specific Plug-in modules. WebGL can work on Video Card which supports OpenGL 2.0 and it is one the standard specification of the latest 3D-CG for the major browsers.
- Implementing the following facilities;
  - reading such data expressed in Excel
  - translating to JSON data by PHP,
  - changing shape of the relevant blocks, and
  - moving viewpoint to watch the map by means of mouse drugging.
- Employing the following JavaScript libraries for the sake of compact/smart implementation;
  - adopting "OBJ" of general 3D file format which is useful for major rendering software for 3D modeling.
  - three.js, which is a powerful library to perform roll-in OBJ-based data and support mouse-based manipulation such as rotation, pin-in/out, zoom-in/out together with WebGL.
  - comparing Away3D TypeScript, but it is not employed.

System implementation is as follows. Figure5 and 6 shows our process of implementation;

1. production of 3D-data as OBJ style : We draw the map of Shikoku with Adobe Illustrator in Figure5.

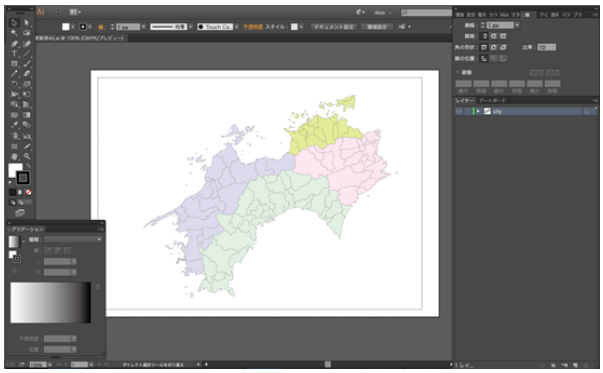


Figure 5. Production of 3D data with "Illustrator".

2. Modification of 3D-data for adjustment:

   Figure6 shows modification of 3D-data in OBJ and translation into OBJ file of 3D-data with "LightWave" for displaying on the Web browsers.

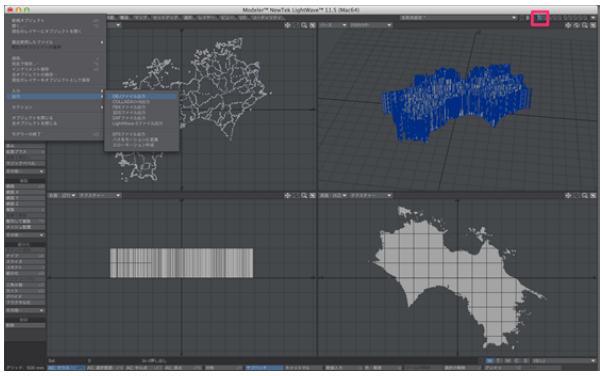


Figure 6. Handling/Modifying of 3D-data for display on WebGL with "LightWave".

ISBN: 978-1-941968-42-0 ©2017 SDIWC

3.3 Visual Design of the 3D map with open accessible data

Figure7 is a trial generation of 3D map without any open data. But we think it is more useful for people to recognize 3D map rather than only 2D one show in Figure4.

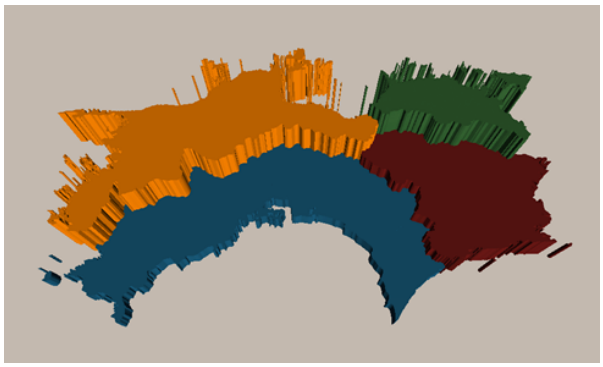


Figure 7. Generated 3D map from Figures3.

Figure8 shows one of the results our system have been able to produce 3D map with open accessible data of population give in Figure3.

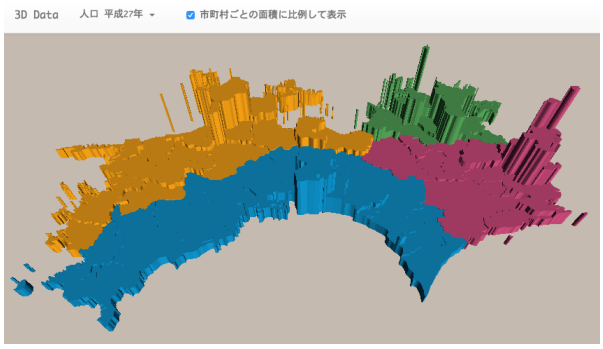


Figure 8. 3D map with open data of population.

4 APPLICATION OF VISUAL DESIGN

Using 3D map with open accessible data, we can very easily obtain flexible viewpoint moving for our suitable application. For example, Figure9 will provide right-hand viewpoint of Figure8.

With around viewing 3D map, we can obtain normally hidden geometric positioning between user specified points and among focused points. For example, we think that application field of our 3D map and visual design with open accessible data will be toward city/urban planing at the local government, more realistic understanding of geography at school, suitable decision-making at Agriculture, Trafic control, Medical problems and so on.

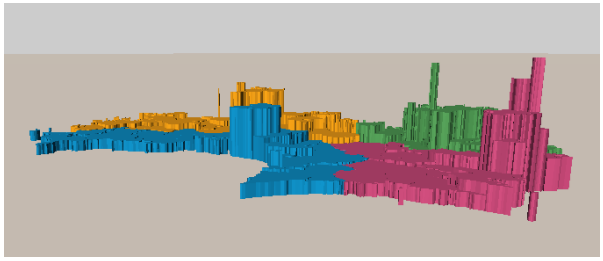


Figure 9. Around viewing for 3D map.

5 CONCLUSION

We have developed our system to produce around viewing facility based on 3D map with open accessible data from the local government. With JavaScript, WebGL and other utilities such LightWave, useful 3D map can be brushed up into more useful and visualizing tool for people to recognize realistic image for their feasible understanding. We will try to apply our system into other domain near future.

Acknowledgement: The authors are thankful to Shikoku Information Communication Conference for their kind contribution to this study.

REFERENCES

[1] C. Kawanishi, et al., "Development and Evaluation of Learner-centric Graphical Educational Tool for Network Study," Proc. of 2013 International Conference on Humanized Systems (ICHS2013@Takamatsu), pp.108–113, Sept. 2013.

[2] K. Higashikakiuchi, et al., "Design, Implementation and Trial Evaluation of a Visual Computer Simulator by JavaScript," Proc. of The Second International Conference on Electronics and Software Science (ICESS2016@ Takamatsu), pp.158–164, Nov. 2016.

[3] `https://medium.com/re-form/nycs-housing-cost-myth-9dce6052c139#.symj85fwt` (accessed on the 20th of July, 2017).

WappenLiteDocker

– A Interface Program between a Web-Browser and a Docker Engine

Koji Kagawa, Haruhiko Nishina and Yoshiro Imai
Kagawa University
2217-20 Hayashi-cho, Takamatsu, Kagawa 761-0396, JAPAN
kagawa@eng.kagawa-u.ac.jp

ABSTRACT

This paper reports WappenLiteDocker which acts as a filter between a Web browser and a Docker engine. It is used to provide a Web-based programming environments to novices and is intended to run on the client machine. This paper reports its features and structures.

KEYWORDS

Programming, Programming Language, Web, Docker, Java,

1 INTRODUCTION

Web-based programming environments are effective for novice learners to learn relatively minor programming languages. They cannot spare so much time in learning minor languages. To provide Web-based programming environments, we need a back-end program that compiles and runs programs while communicating with Web-browsers, unless we can provide JavaScript-based implementations. There are several existing server-side platforms that provide such environments. For example, an online programming environment Paiza.io (https://paiza.io/) provides server-side environments to execute programming language implementations using Docker (https://www.docker.com/). Docker is a software container platform to allow programs run in an isolated setting without heavy overhead and is therefore suitable for providing back-end programs for Web-based programming environments. However, in such server-side environments, teachers cannot freely prepare programming language environments and customize such environments as they like. And it is very unlikely for service providers to give teachers such freedom because it would be very difficult to guarantee security of containers. Running server-side programs sometimes makes the server overloaded when the server has hundreds of clients.

One of the authors has proposed WappenLite [1] for Java Virtual Machine (JVM)-based languages for providing Web-based programming environments. WappenLite is implemented in Java and has a feature that the back-end program runs in the client machine instead of the server machine. WappenLite can run JVM-based language implementations in a sandbox and has been effective for JVM-based languages. It is, however, helpless for non-JVM languages. Though it is possible to invoke native commands via the `java.lang.Process` class, it cannot provide sandboxes in such cases. Moreover, it is delicate to the change of environments caused e.g. by a version up of compiler toolkits. For example, there is no popular JVM-based implementation for C nor Haskell. A C compiler is also necessary for running special purpose languages such as Flex and Bison which are used in the course of compilers.

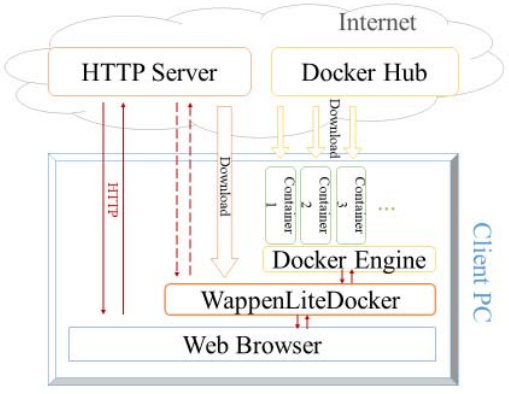


Figure 1. Client-side Mode of WappenLiteDocker

In this paper, we will propose a program called WappenLiteDocker which uses Docker on the client-side machine. We think that it is not unrealistic to assume that students in computer science departments have already installed Docker (Docker for Windows/Mac or Docker Toolbox) in their PCs. WappenLiteDocker acts as an interface between a Web browser and a Docker engine which resides in the same computer as the Web browser (Figure 1).

Running back-end programs on the client-side computer has several advantages. First, teachers do not have to prepare a dedicated server. Second, it is possible to introduce broader range of example programs that deal with multimedia (i.e. large-size) data such as images and sounds and possibly graphical user interface (GUI). Finally, it facilitates additional processing of user programs which might be too heavy to execute on a single server hardware. For example, some kinds of program transformation might be necessary to introduce program visualization of user programs (e.g. [2]).

We use Java as an implementation language for our program. It is chosen due to its platform independence and rich network-related libraries. Another reasonable option would be to use the Go language which is used for implementing the Docker command-line programs. We choose Java simply because the authors are more familiar with Java than Go and because the potential users (teachers) of our system are also more likely to be familiar with Java.

The structure of this paper is organized as follows. First, we will explain the requirement to our system in Sect. 2. Then, Sect. 3 will explain the internal structure of the system. Section 4 will discuss future directions and concludes.

2 REQUIREMENTS

In this section, we will explain which are required and which are not required for our program.

2.1 CORS

Docker engines already have an HTTP-based interface called Docker Engine API (https://docs.docker.com/engine/api/). Therefore, it is possible to talk to the Docker engine using Web-browsers and HTTP client programs such as `wget` and `curl`. Why do we need an interface program such as WappenLiteDocker? This is partly because of the restriction of cross-origin requests. Though users can directly type the URL provided by the Docker engine such as http://127.0.0.1:2375/containers/json to connect to the Docker engine, pages downloaded from another site, say http://www.example.com/, cannot access the above Docker URL using JavaScript. This restriction can be relaxed by attaching Cross Origin Resource Sharing (CORS) headers to the response of the Docker API. This could be done by using a reverse proxy program such as Nginx or adding the `--api-cors-header` option when starting the Docker daemon, if this were the only necessary action.

2.2 Security

Unfortunately, this approach is vulnerable to malicious Web pages especially when teachers cannot prepare a dedicated HTTP server for his Web pages. Malicious users would be able to devise Web pages that can create a Docker container from any Docker image with any create options and put the prepared pages in the HTTP server. Therefore, we need to restrict the Docker API commands which are passed to the Docker engine. Our filter program passes the URLs shown in Table 1 only, which shows, in this first column, paths after `/v`*`n.nn`*`/container` where *`n.nn`* is the version number of API and *`id`* is a container identity number. Then, Web pages that communicate with the Docker engine via WappenLiteDocker can do only the following actions. First, they can list containers, create start, stop, restart and delete containers. These are necessary to manage containers. Second, they can send files to containers and receive files from containers. These are necessary be-

cause we want to manipulate programs that deal with multimedia data such as images and sounds. Finally, they can attach to the standard input and output of containers via WebSockets. This is necessary because user programs may deal with large data and therefore they may not finish immediately after starting. Then, it would be necessary to push data from the container to the browser later.

Table 1. URLs which are allowed to pass

Path	Method	Remark
`/json`	GET	
`/create`	POST	See Table 2
`/id/start`	POST	
`/id/restart`	POST	
`/id/stop`	POST	
`/id`	DELETE	
`/id/attach/ws`	GET	WebSocket
`/id/archive`	GET	
`/id/archive`	PUT	

We further restrict the parameters of the command for creating containers as in Table 2.

Table 2. Parameters of `/create`

Parameter	Restriction
`Image`	Only image names which are hard-coded in the source code are allowed.
`ExposedPorts`	Not allowed.
`Volumes`	Not allowed.

In order to prevent malicious users from invoking dangerous images, we require that valid (i.e. considered to be safe) image names should be hard-coded in the source code of WappenLiteDocker. This is a strong requirement and this is why the choice of implementation language is important to our system. If it were written in an unfamiliar language, teachers would hesitate to rewrite the system. Moreover, `ExposedPorts` and `Volumes` are not allowed in the parameter. The former is used to expose TCP/UDP ports in the container and the latter is used to mount the file system of the host computer in the container. These parameters are known to be often exploited to hijack the host computer.

2.3 Non-requirements

Since we run programs on the client side, we do not have to care much about the learners' mistake such as nonterminating programs. Learners can stop the runaway container by other means since it runs on their own computers. Though it would be possible to be more user-friendly in this respect in the future, the current version does not pay much attention to troubles caused by learners' mistake.

3 STRUCTURE

To realize the above requirement, we can use rich network-related libraries provided in Java. It is implemented as an Eclipse project using Maven and the program is distributed as a runnable JAR file which packs all the necessary libraries in a single file.

We use the Async Http client library (https://github.com/AsyncHttpClient) to access the Docker Engine API and use the Jetty library (http://www.eclipse.org/jetty/) to act as the HTTP server and to wait HTTP connections from browsers. Some Docker environments such as the Docker Toolbox requires client-side authentication to access the Docker API. We use Bouncy Castle (https://www.bouncycastle.org/), which is a cryptography library for Java, for this purpose.

Internally, WappenLiteDocker currently consists of six Java classes. Among them, two are servlet classes and one is a filter class. The filter class is used for adding CORS headers to HTML responses. One servlet class is devoted to handle WebSocket requests and the other servlet class does most of the necessary work of WappenLiteDocker. For example, it checks the parameters of the create command and filters out the forbidden parameters. Moreover, it registers the identity numbers of created containers so that it only passes commands targeted at the containers created by WappenLiteDocker. The rest of classes consists of one utility class for creating HTTPS connections and two classes that contains the main method.

5 CONCLUSION AND FUTURE WORK

Currently, WappenLiteDocker is implemented as is described in the previous section and is tested for two Docker images `gcc` and `haskell`. It is not put into practical use yet, that is, it is not used in actual classes. We plan to use it in the class for introductory C programming with a program visualizer presented in [2]. When we tried to deploy the program transformer used in the above program visualizer in our server machine, it did not work well because the program transformer was too heavy to run on a single computer for nearly a hundred of clients. We expect that we can circumvent this problem by running the program transformer on the client PCs.

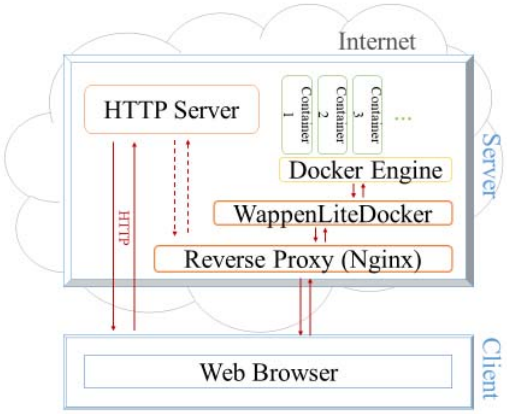


Figure 2. Server-side Mode of WappenLiteDocker

In the current version, WappenLiteDocker simply passes HTTP request unmodified to the Docker engine. However, the raw Docker engine API is somewhat cumbersome to use from the viewpoint of JavaScript programs. For example, we have to manipulate the TAR file format in JavaScript. Therefore, we want make the API simpler and handier to use from JavaScript programs.

Though the primary target of WappenLiteDocker is to use it on client PCs, it would be also useful if it can run on a server machine (Figure 2). It is inevitable to use the server-side mode, when learners use platforms such as iOS and Android simply because Docker does not support such platforms. We would like to add features to support the server-side mode in the future version of WappenLiteDocker. For example, it would be necessary to add options to limit container resources such as CPUs and memories forcibly. Though it does not have a user authentication mechanism, it can use a reverse proxy program such as Nginx in front and use the basic HTTP authentication mechanism provided by Nginx. Though there are several similar interface programs such as `io.livecode.ch` that connect Docker with browsers, a Java-based platform would be beneficial to a certain amount of teachers.

Since it allows teachers to use arbitrary containers they want, it would not be practical to ensure the security of the host computer perfectly. Instead, we would like to keep the program simple and to suggest teachers to use it in a restricted setting, for example, when it is accessible only within the campus network or accessible only during a week before the deadline of a report submission.

Acknowledgements

This work is partially supported by JSPS KAKENHI Grant Number 15K01075.

We would like to thank Keita Kimura for his contribution to the early prototype of the system.

REFERENCES

[1] K. Kagawa, “WappenLite: a Web Application Framework for Lightweight Programming Environments,” 9th International Conference on Information Technology Based Higher Education and Training (ITHET 2010) pp. 21-26, April 2010, Cappadocia, Turkey.

[2] T. Suetomo and K. Kagawa, “A Program Visualization System using 3D Still Graphics,” 14th International Conference on Information Technology Based Higher Education and Training (ITHET 2015), 4 pages, June 2015, Lisbon, Portugal.

Arduino Based Automatic Guitar Tuning System

Matej Žerjav
University of Ljubljana
Faculty of Mechanical Engineering
Askerceva 6
1000 Ljubljana, Slovenia
Email: zerjav.matej@gmail.com

Primož Podržaj
University of Ljubljana
Faculty of Mechanical Engineering
Askerceva 6
1000 Ljubljana, Slovenia
Email: primoz.podrzaj@fs.uni-lj.si

***Abstract*—Acoustic guitar is one of the best known and most commonly used musical instruments. It emits sound waves because the guitar player makes the strings oscillate. In order to function properly, i.e. emit the sound of the appropriate frequency it is very important the tension in each string is adjusted regularly. This process is called tuning. It is usually done manually with a help of universal tuner. In this paper an automatic guitar tuning system is presented. It is based on the Arduino platform. It uses a microphone to record the sound, calculates its frequency and then turns the tuning pegs on guitar head with stepper motor to get the correct frequency.**

***Keywords*—IEEE, IEEEtran, journal, LaTeX, paper, template.**

I. Introduction

Many engineers working in the field of Mechatronics say, that they gained the most knowledge, when they had to realize some real life application. Although the underlying theory from both Mechanics and Electronics, which give the name Mechatronics, when combined, is of course important for the understanding of key concepts, real life applications give a chance to students to memorize these concepts easier. For this reason it is important to widen the set of mechatronic applications as much as possible, so that everyone can see its applicability and also to make it possible to find an application that a specific person is most familiar with.

In recent years Arduino platform has become one of the most popular (if not the most popular) platforms for real time applications, because it is very inexpensive and quite easy to use. Applications range from drones [1] and robot dogs [2] to motor control [3] and portable solar power source [4]. In this paper an application from the field of acoustics will be presented. It is well known that guitar as all the other stringed instruments has to be tuned in order to produce the sound of the correct frequency. Usually this process is done manually. In this paper an Arduino based automatic guitar tuning system will be presented.

II. Theoretical background

A. String oscillations and pressure waves in gas

It is well known that in physics sound can be described as the oscillations in gas pressure Δp measured from some equilibrium value [5]:

$$\Delta p = \Delta p_{max} \sin(kx - \omega t) \tag{1}$$

The equation represents longitudinal waves in space with the angular wave number k and the angular frequency ω. How musical instruments achieve to produce this waves depends on the type of the instrument. In the case of stringed instruments, this oscillations of pressure are result of the oscillation of various strings. In order to get the relation between the string motion and air pressure we first need the relation between the parameters of the string and the speed of waves on a string. It is given by the following equation [5]:

$$v = \sqrt{\frac{T}{\mu}} \tag{2}$$

where T is the tension in the string and μ the mass per unit length of the string. If the string is fixed at both ends (see the first plot in Fig. 1) the phenomenons of reflection and interference have to be taken into account. This results in the occurrence of standing waves. Different modes of oscillations are possible (see bottom three plots in Fig. 1). The points of zero amplitude are called nodes and the points of maximum amplitude are called antinodes. The section of the standing

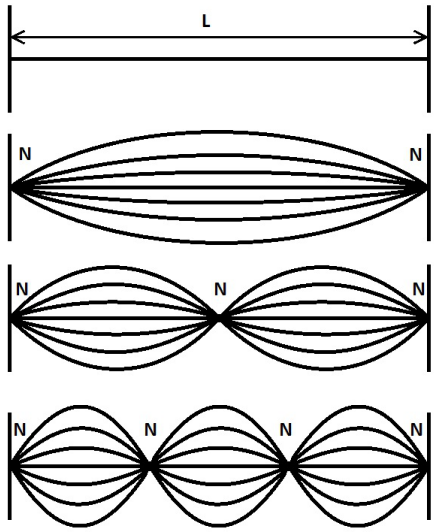


Fig. 1. Schematic representation of a guitar string oscillating with a fundamental frequency (second plot) and first and second harmonic (third and fourth plot)

wave from one node to the next node is called a loop. The frequency of oscillation is given by the following equation [5]:

$$f_n = \frac{n}{2L}\sqrt{\frac{T}{\mu}} \quad n = 1, 2, 3, \ldots \tag{3}$$

where L is the length of the string and n the number of loops. The lowest frequency (associated with the second plot in Fig. 1), when $n = 1$ is called the fundamental frequency and is therefore given by the following equation:

$$f_1 = \frac{1}{2L}\sqrt{\frac{T}{\mu}} \tag{4}$$

Frequencies for $n \geq 2$ are called harmonics.

B. Acoustic guitar

An acoustic guitar (see Fig. 2) is an instrument with six strings. Its main part is the body to which a neck is attached.

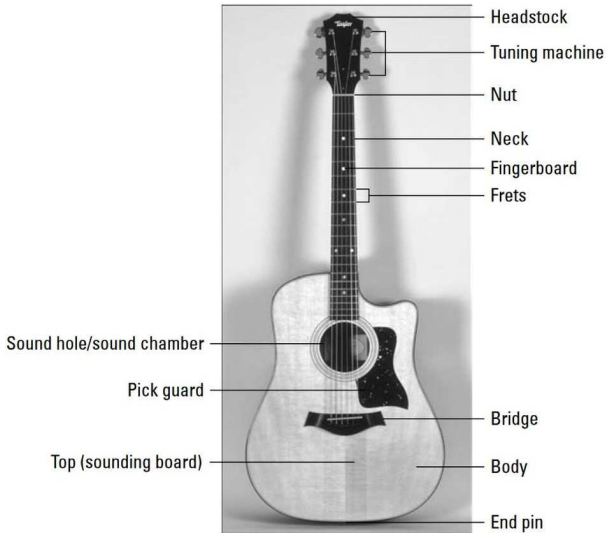


Fig. 2. Acoustic guitar and its basic components [6]

The strings are attached to the bridge on the body and to the head at the end of the neck. The strings differ in mass per length (parameter μ in Eq. 4. The parameter L can be modified during playing by the position of the fingering point. The plucking point determines the presence (the amplitudes) of the higher harmonics. When a string is plucked close to the bridge a tone that is softer in volume is produced. It is also brighter and sharper, and the sound is richer in high frequency components [7]. The other extreme case is when plucking is done near or over the fingerboard, closer to the midpoint of the string. In this case, the tone is louder, mellower and less rich in high frequency components. The fundamental frequencies for each of the strings are given in Table I. The exact frequency can be obtained by a proper adjustment of the string tension (parameter T in Eq. 4. For this purpose each acoustic guitar has the so called tuning machine located on the head of the guitar. The appropriate tension can be obtained by

TABLE I
THE FUNDAMENTAL FREQUENCIES OF GUITAR STRINGS [8]

String	Note	Octave	Frequency [Hz]
6	E	2	82,41
5	A	2	110,00
4	D	3	146,83
3	G	3	196,00
2	H	3	246,94
1	E	4	329,63

the adjustment of the tuning pegs. This process is called the (guitar) tuning.

As humans are unable to exactly determine the frequency of the audible sound, some kind of device to assist in this process. The most common one is the universal tuner shown in Fig. 3. The user interface makes it possible to specify

Fig. 3. Fire&Stone / Coxx CLIP-UT Tuner

the string we want to tune. The tuner then measures the (fundamental) frequency of the sound emitted by the guitar and gives information about the appropriate movement of the specific tuning peg. After several iterations the string is tuned within a specific range of the frequency given in Table. I. Of course this process has to be repeated for each string.

III. EXPERIMENTAL SETUP

The main building block of an automated guitar tuning system is a controller in which control algorithm programmed. For our application, we have decided for Arduino Zero (see Fig. 4). It is namely very inexpensive, commonly used and

Fig. 4. Arduino Zero

user friendly system. The development environment that can be installed free of charge includes many examples. These can then be easily combined to perform more complex control related tasks. In our case the main tasks are the acquisition of the fundamental frequency in real time and the drive of a stepper motor, which turns the tuning peg on the guitar

head. The scheme of the whole system is given in Fig. 5. In the lower left corner of the electret microphone is shown.

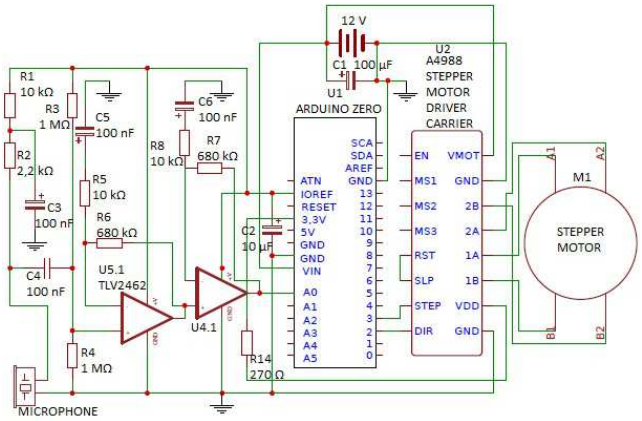


Fig. 5. The electrical scheme of the whole system

Its output signal is then amplified twice before it is fed into the Arduino Zero. This is achieved by the TLV2462 Dual Operational Amplifier shown in Fig. 6. The relation between

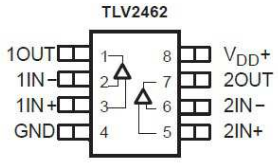


Fig. 6. TLV2462 Dual Operational Amplifier

output voltage V_{out} and two input voltages (V_1 and V_2) is given by the following formula [9], [10]:

$$V_{out} = G_V(V_2 - V_1) \tag{5}$$

where G_V is the voltage gain. It can be controlled by the values of the resistors in input and feedback paths [11]. After the resulting signal is obtained by the Arduino, it is very easy to get the frequency. The code needed to get is namely very simple (see Fig. 7). When the actual fundamental frequency

```
int frequency = meter.getFrequency();
  if (frequency > 0){
    Serial.print(frequency);
    Serial.print(" Hz    ");
```

Fig. 7. The part of code used to get (and write) the fundamental frequency.

is obtained, we just need to compare it with the desired fundamental frequency for the specific string. If the actual fundamental frequency is within a certain predefined range around the desired one nothing needs to be done. If it is either too low or too high, the tuning peg must be turned in the correct direction. In order to be able to turn the tuning pegs on the guitar head a special part had to be made (see Fig. 8).

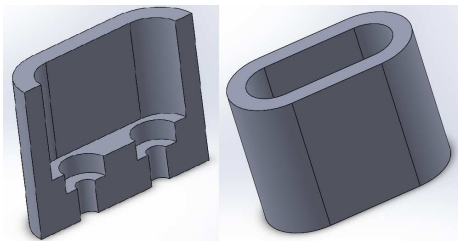


Fig. 8. A part of the system used to turn the tuning pegs

As it must fit the selected stepper motor on one side and the tuning peg on the other, it is difficult to find it. For this reason it was the only part that had to be designed specifically for this application. This was done in SolidWorks and produced using a 3D printer. The part attached to the stepper motor is shown in Fig 9. The stepper motor used was a NEMA 17 stepper motor

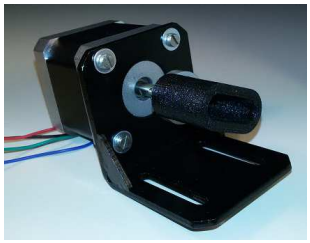


Fig. 9. Stepper motor

(type 42HS4013A4 with a $1.8°$ step angle). In order to drive it a A4988 Stepper Motor Driver was used. The part of the code used to drive the stepper motor is shown in Fig. 10. The

```
digitalWrite(dirPin,LOW);
for(int x=0; x<15 ; x++ ) {
  digitalWrite(stepPin,HIGH);
  delayMicroseconds(500);
  digitalWrite(stepPin,LOW);
  delayMicroseconds(500);
    }
  delay(2000);
```

Fig. 10. The part of the code used to drive the stepper motor

last delay in the code was selected to be 2 seconds in order to hear and sense the change in frequency of the emitted sound. Some delay is needed in any case, because the system needs some time to determine the frequency correctly.

The stepper motor itself than has to be fixed relative to the guitar head. A special fixture was made for that purpose. The setup is shown in Fig 11. The only part shown in this figure that was not mentioned before is the battery at the top of the figure. The guitar player now only has to put the very light stepper motor into the position and specify which string he wants to tune. After that the string is plucked and the system autonomously tunes the string in.

ISBN: 978-1-941968-42-0 ©2017 SDIWC

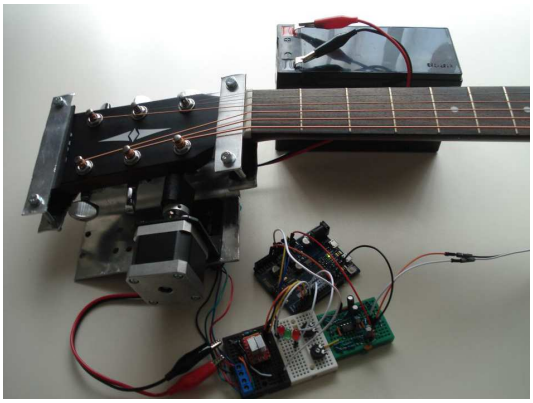


Fig. 11. System setup

IV. Results

In order to verify the precision of the presented automated guitar tuning system it was used in connection with the universal tuner shown in Fig. 3. The only purpose of the tuner is to check if the correct fundamental frequency is obtained. The whole system used for verification is shown in Fig. 12. When tuning is done manually the universal tuner screen is

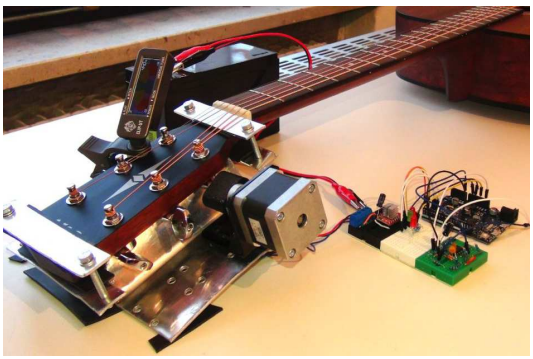


Fig. 12. The photo of the whole system used for verification

observed. Fig. 13 shows the display for two special cases. The

Fig. 13. Universal tuner display (upper photo when frequency is too low; lower photo when frequency is appropriate

upper photo in Fig. 13 shows the display when the frequency is too low. The guitar player must increase the tension in the string. This is done by tuning peg rotation. This step is repeated iteratively until the universal tuner signals that and appropriate frequency has been achieved (lower photo in Fig. 13).

When tuning is done automatically, the guitar player first has to select which string (actually which note to be more exact), he wants to tune. Then the stepper motor is put into the correct position (see Fig. 12) and the string is plucked. The system is then able to autonomously turn the tuning peg into the correct position. There are only two requirements that have to be fulfilled:

- The starting tension in the string must not be too far off the appropriate value. If the string is much too loose the system has difficulties to when trying to determine the fundamental frequency of oscillations.
- Even if the tension of the string is not too far off the appropriate value, it may be needed to pluck the string several times. Especially if we do it very gently. The amplitude of oscillations namely steadily decreases. In order that the system can function, the amplitude of oscillation must be above a certain threshold. It should be noted that how many times the string should be plucked also depends on the selected time delay in the last line of the code given in Fig. 10.

Several guitar players were asked to try the system. After it was explained to them, how the system functions, all of them were able to tune the guitar easily using the presented system.

V. Conclusion

The main purpose of this paper is to present another possible mechatronic application for Arduino based platforms. Guitar tuning is of course done by guitar players, who of course are more artists in nature and often not too interested in science and engineering. The presented automatic guitar tuning system offers them an interesting first step into the field of engineering and electronics, which they might make.

In current setup the system of course lacks in applicability. As we used the components that were used in other projects, many of them are not optimised. Battery for example much too large. If an optimisation were made (especially in the case of battery and in the case of the fixing of the stepper motor, the system would also be of practical value as some of the guitar player that we had invited to test the system have noted.

References

[1] D. McGriffy, *Make: Drones: Teach an Arduino to Fly*. Maker Media, Inc., 2016.

[2] A. J. Boloor, *Arduino by Example*. Packt Publishing Ltd, 2015.

[3] J. Boxall, *Arduino workshop: A Hands-On introduction with 65 projects*. No Starch Press, 2013.

[4] J. Purdum and D. Kidder, *Arduino Projects for Amateur Radio*. McGraw Hill, 2015.

[5] R. A. Serway and J. W.Jewett, *Physics for scientists and engineers with modern physics*, 9th ed. Brooks/Cole, 2014.

[6] M. Phillips and J. Chappell, *Guitar for Dummies, 2nd edition*. Indianapolis, Indiana: Wiley Publishing, 2006.

[7] C. Traube and J. O. Smith, "Estimating the plucking point on a guitar string," in *Proceedings of the COST G-6 Conference on Digital Audio Effects, Verona, Italy*, 2000.

[8] W. M. Hartmann, *Principles of Musical Acoustics*. New York: Springer, 2013.

[9] P. Horowitz and W. Hill, *The art of electronics*, 3rd ed. Cambridge University Press, 2015.

[10] S. T. Karris, *Electronic Devices and Amplifier Circuits with MATLAB Computing*, 2nd ed. Orchard Publications, 2008.

[11] R. L. Boylestad and L. Nashelsky, *Electronic Devices and Circuit Theory*. Pearson Education, 2013.

Development of Document Transferring and Archiving Service with Sentiment Analysis-based Preprocessing Facility

Shunsuke Doi* Yoshiro Imai Kazuaki Ando Koji Kagawa
Rihito Yaegashi Keizo Saisho Kyosuke Takahashi Hitoshi Inomo
Naka Gotoda Toshihiro Hayashi Hiroyuki Tominaga Tomohiko Takagi
Graduate School of Engineering, Kagawa University
2217-20 Hayashi-cho, Takamatsu, Kagawa pref., Japan
s17g471@stu.kagawa-u.ac.jp*,
{imai,kagawa,ando,rihito,sai,k_taka,inomo,gotoda,hayashi,tominaga,takagi}@eng.kagawa-u.ac.jp

ABSTRACT

People used to be writing several kinds of documents such as memoranda, message, e-mail etc. for the third persons to read possibly with emotional feeling. After they prepared the above documents, sometimes such documents might unintentionally hurt other's heart due to our careless emotional expressions. If some kind of checking services were utilized for the above careless emotional expressions, people could avoid to write documents which would unintentionally hurt other's heart. This paper describes our newly developed Document Transferring and Archiving Service with Sentiment Analysis-based Preprocessing Facility. It is realized as server-client computing model, namely its server is written in Perl and PHP just like LAMP(Linux-Apache-MySQL-PHP/Perl) and its client is written in JavaScript executing on the major Browsers. The service can scan the regarding document of a user, separate it into word-level expressions, check them against sentiment dictionary, calculate each sentimental values for document and generate the corresponding radar chart for the document based on emotional axises such as delight, anger, sorrow, pleasure and so on. Users, namely writers, use our Service before transferring and/or archiving, they can check their documents by means of the above preprocessing facility and recognize how their ones have a lot of emotional feelings which would include non-suitably emotional expressions. With such a service, users of the service can avoid to write, transfer or archive such documents which would unfortunately make someone feel bad.

KEYWORDS

Sentimental value, Sentiment analysis, Visualization, Japanese Language processing, Document transferring, Document archiving.

1 INTRODUCTION

Data mining becomes more and more popular and then big data analysis has been consistent with the trend of information processing. Data mining and big data analysis seem to be excellent approaches to reveal hidden relation behind phenomena people have never found before and bring new recognition about such a relation to our daily lives for possible decision-making.

Sentiment analysis[1] can play another important role of information processing for sentences, expressions and documents not only from SNS/Internet but also from normal, daily verbal communication. Due to preprocess related with sentimental analysis, we can find whether our writing document includes "unnecessary or non-suitable" emotional expressions or not. So we would be relieved to transfer our documents as well as archive ones if we had employed any preprocess to detect such emotional expressions and then had removed these expressions.

This paper describes our Document Transferring and Archiving Service with Sentiment Analysis-based Preprocessing Facility. Such service is designed and implemented as server-client computing model and for users to manipulate it through Web-based application. The paper introduces a whole system and then explains a preprocess service to generate Senti-

mental Dictionary and to perform Sentimental Analysis to detect emotional expressions graphically. It illustrates some trial of visualization of documents through sentimental analysis in the third section. It demonstrates document transferring and archiving by means of our services in the fourth section. And finally it concludes our summaries in the last section.

2 SYSTEM CONFIGURATION AND ITS PREPARATION PROCESS

This section explains our system with its main flow and its important preprocess to define sentimental dictionary for preprocessing facilities of sentimental analysis.

2.1 Main Flow of System

Our system has been design and implemented as a typical server-client computing model, where the server is realized in the LAMP type and the client executes on the major browsers written by JavaScript. Figure1 shows a main flow of the system, the three major menus, namely input menu, show graph one and e-mail transfer one, are provided on the client and then the related three procedures, namely sentimental analyzing, simple mail transferring and PDF generating are invoked on the server and performed as server tasks written in PHP and Perl scripts. The system has two major database managing modules which execute on the server together with SQL database.

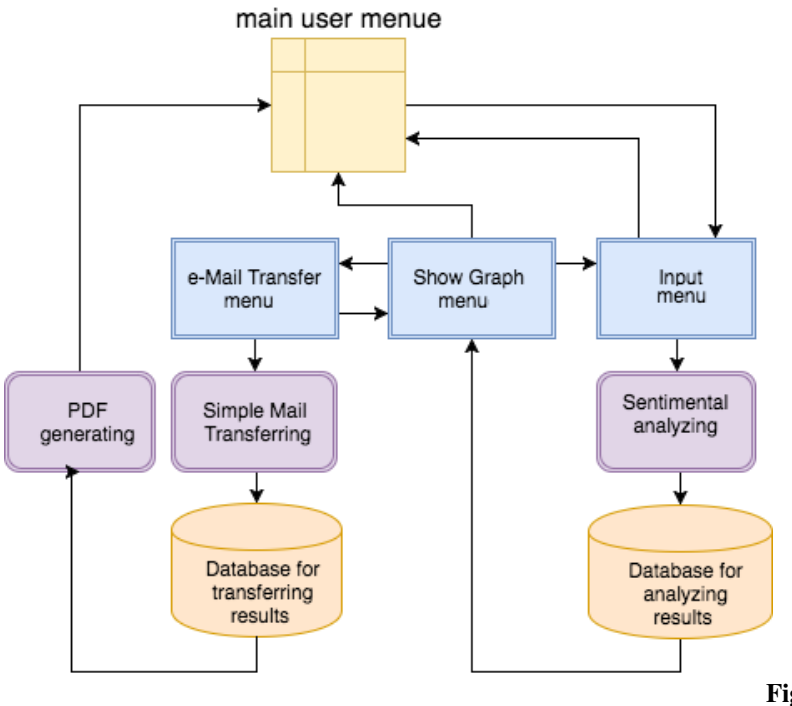


Figure 1. Main flow of the system

2.2 Definition of sentimental dictionary for preprocess

Before describing our sentimental analysis for preprocessing, this subsection explains how to define sentimental dictionary and how to calculate sentimental value for each focused sentence. With Japanese language processing, our system obtains some documents in the target domain, decomposes such a document into sentences and decompose them into a series of words by MeCab. Such words are classi-

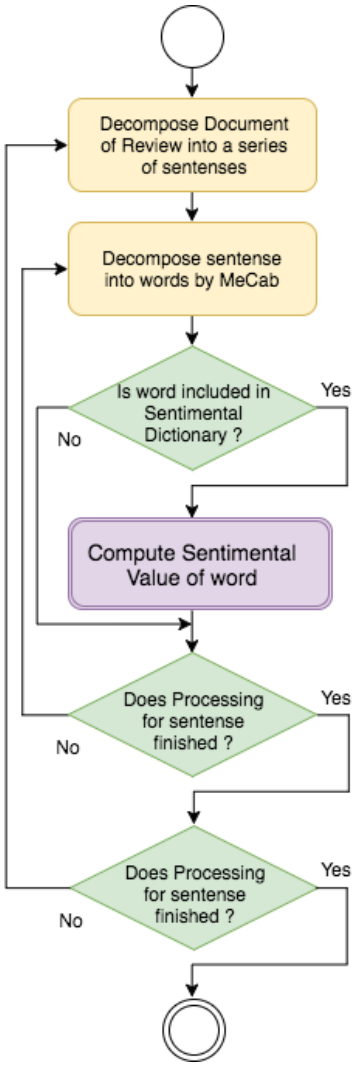


Figure 2. Flow to define the dedicated sentimental dictionary for Preprocessing

fied into cluster and computed with sentimental values according to sentimental dictionary[2]. Figure2 shows the above procedure to define applicable sentimental dictionary for the succeeding sentimental analysis.

3 VISUALIZATION THROUGH SENTIMENTAL ANALYSIS

This section demonstrates visualization of sentimental value for input document from user client by means of sentimental analysis[3]. It shows process flow for sentimental analysis at first, illustrates how to calculate sentimental values for each sentence, for example in case of negation and conjunction, and demonstrates visualization of sentimental values of document by means of Graph charts generating module of Web-based client.

3.1 Process flow for sentimental analysis

Process for Sentimental Analysis starts decomposing the obtained document into sentence, decomposing each sentence into a series of words by MeCab, and then checks whether each word is registered in the sentimental dictionary or not. If each word of the sentence is registered, then sentence is calculated with sentimental values according to the conditions including “Negation” and/or “Conjunction”. Because Negation and/or Conjunction is included in the sentence, calculation of whole sentimental value for the relevant sentence must be modified in the ways shown in Figure3 and Figure4 . Figure5 shows process flow for sentimental analysis.

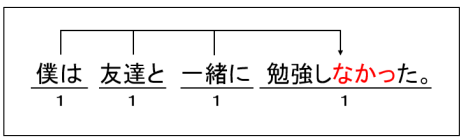


Figure 3. Example for Determination of Negation

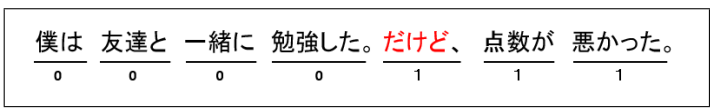


Figure 4. Example for Determination of Conjunction

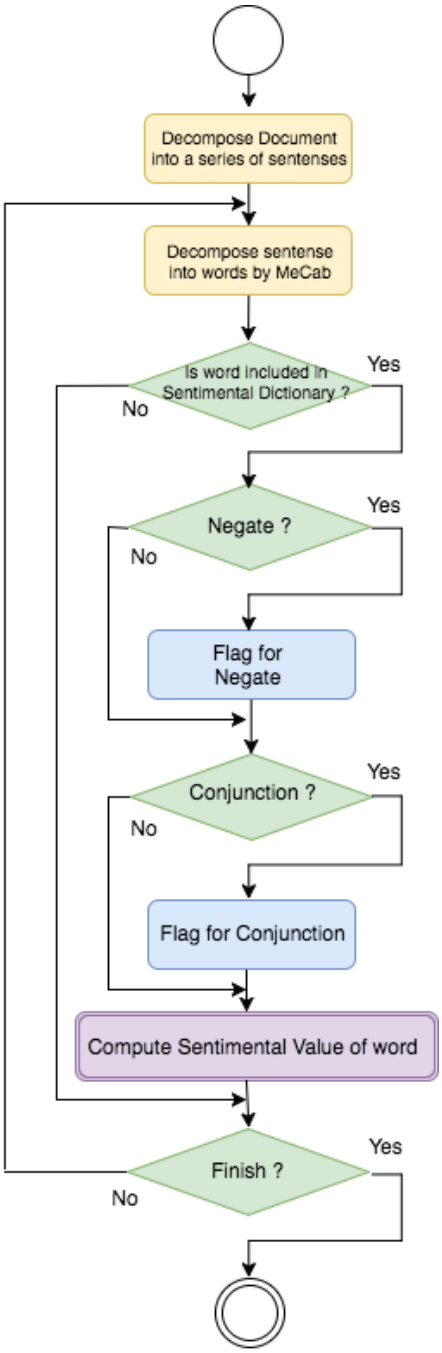


Figure 5. Process Flow for Sentimental Analysis.

3.2 Computing and assigning of sentimental values

After picking up the relevant sentimental value for each word from sentimental dictionary together with detection of “Negation” and “Conjunction”, calculation of sentimental values for a whole sentence can be performed in the way shown in Figure5. Real example of calculation of sentimental value for simple document is illustrated in Figure6. A lots of emotional feel-

ISBN: 978-1-941968-42-0 ©2017 SDIWC

ings have been classified and categorized into 10 clusters, for example, ‘delight’, ‘anger’, ‘sorrow’ and ‘pleasure’ are the four major axis, and then we have defined other four clusters, some of them seem to be combined and modified with the four major axis. The bottom two lines of 5 clusters shown in Figure6, which indicate the sentimental values for 10 clusters, are the results of calculated sentimental values for the relevant document described in the expression at the top line of Figure6.

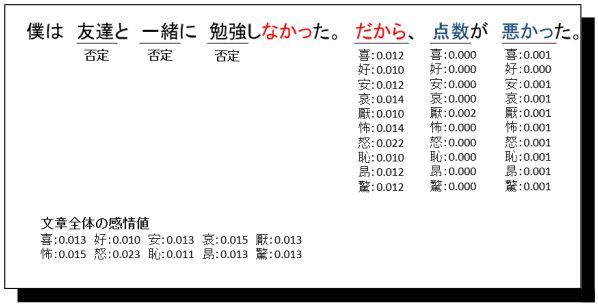


Figure 6. Example of Calculation and Assignment of Sentimental Values for Simple Document.

The system is ready to visualize the results of sentimental analysis for the relevant document as calculated in the way shown in Figure6. This is a preprocess of the system which can provide document transferring and/or archiving service described in the section4.

3.3 Visualization based on the results of Sentimental Analysis No1: from Input to Radar-Chart Generation

Of course, users can understand what kinds of emotional feelings are included in their writing document, for example, by means of recognizing data from Figure6. Therefore, it is more convenient for users to understand emotional feeling whether their writing document includes emotional expressions or not through graphical expression rather than through Figure6. Carefully using with those information, writers, namely users of our system, would be able to avoid such non-suitable emotional expressions thank to visualization of emotional feeling in the relevant document.
As you know, it is more useful for visualization to show the results of sentimental analysis

Figure 7. Input of Document through Web Application : input mode

graphically than to show them in the ways of calculated tabulation or numerical expression just like Figure6. So our system can obtain document by user through input mode shown in Figure7. And then it can calculate sentimental value and visualize the result of sentimental

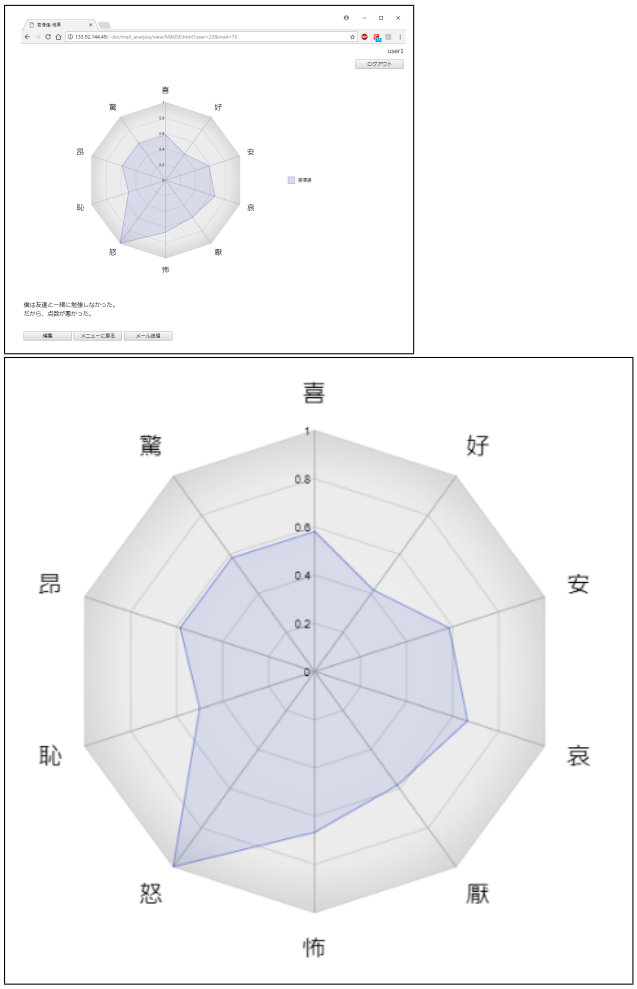


Figure 8. Example for Visualization of Sentimental Values; the upper: visualization of sentimental value on Web app., the lower: generation of radar chart

ISBN: 978-1-941968-42-0 ©2017 SDIWC

analysis graphically just like Figure8. So users can modify their writing documents with emotional feeling into more acceptable contents which includes suitably emotional expressions and reduces non-suitable ones according to visualized results of calculating sentimental analysis for their previous written documents.

3.4 Visualization based on the results of Sentimental Analysis No2: from sequential line graph to mailing of document

User can check how their writing document varies in chronological order from the viewpoint of sequential line graph for the results of calculating sentimental analysis. Figure9 shows sequential line graph for the results of calculating sentimental analysis. It is very useful and attractive for users to understand what kind of emotional expressions in their writing documents vary in chronological order and to obtain a meaningful hint to modify their documents into more suitable not only for themselves but also for those who shall probably read them near future.

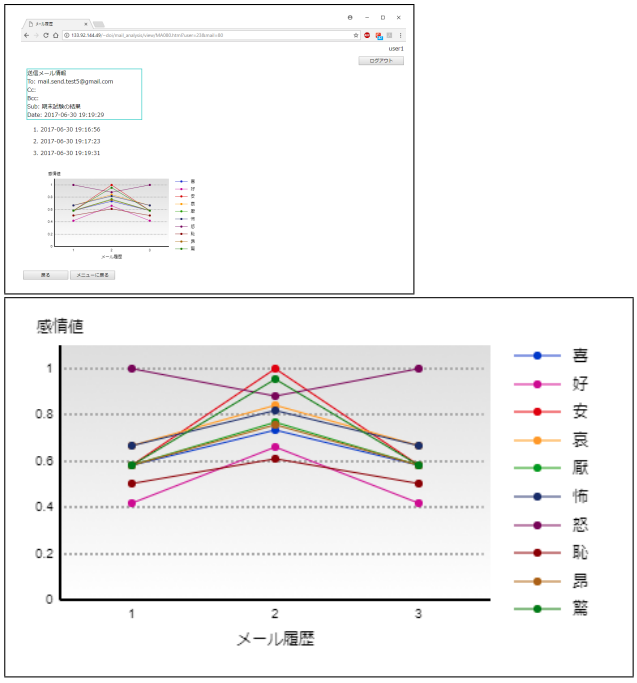


Figure 9. Example for another Visualization of Sentimental Values; the upper: visualization of sentimental value and mailing of the relevant document on Web app., the lower: Transition/Time Series of Sentimental Values

4 DOCUMENT TRANSFERRING AND ARCHIVING SERVICE

This section describes other useful services for document transferring and archiving after sentimental analysis and visualization of its calculated results.

4.1 Document transferring realized by e-mail facility

Document transferring service has been implemented by invocation of server-side e-mail handling facilities called “Postfix” as MTA and “Dovecot” as MUA. The service flow is realized by the way shown in Figure10. Users write their document onto the Web-based client of the system, check what kinds of emotional feelings are included in their writing document, and then instruct to let the relevant document, which is already checked by Sentimental analysis, to be transferred into someone’s e-mail address. Figure11 shows e-mail transfer menu of our system. With our system, the

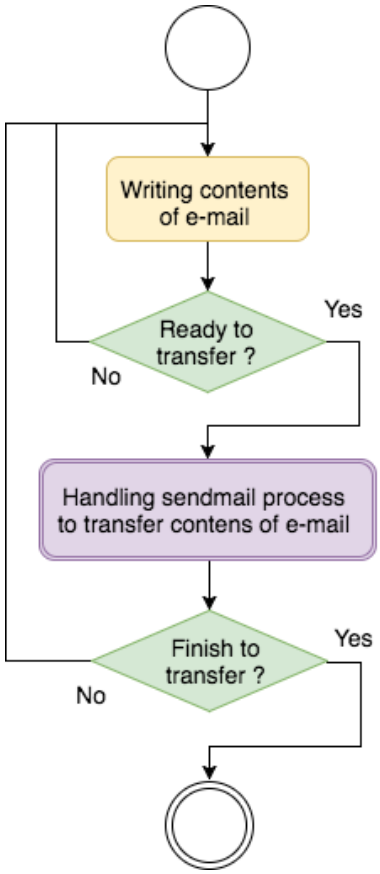


Figure 10. Mail Transferring flow

ISBN: 978-1-941968-42-0 ©2017 SDIWC

document to be transferred shall be modified without non-suitable emotional feelings.

Figure 11. Mail Transferring on Web app.

4.2 Document Archiving by LaTeX and SQL Database

Our archiving service of the system is to generate a PDF style, namely the system can support to transform the documents into PDF. By means of usage of LaTeX, the relevant document can be formed into more easily readable together with dating, signature and other important information automatically. Figure12 shows PDF generating flow for our system which provides to realize archiving service. And Figure13 shows PDF generating menu of our system.

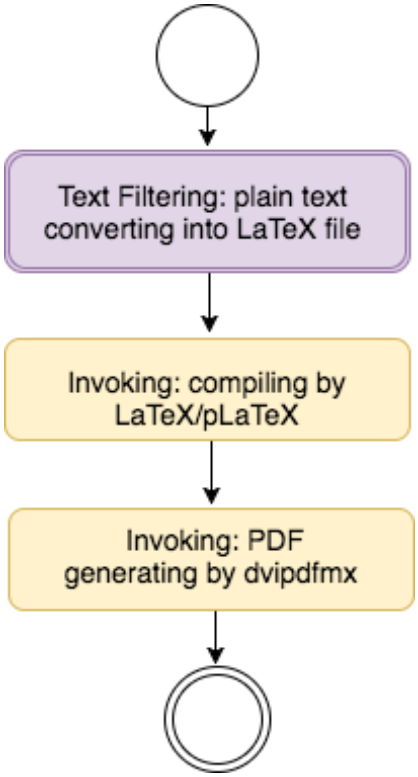


Figure 12. PDF generating by LaTeX

Figure 13. Menu for Generating PDF

With our transferring and archiving service, user would be able to write their documents without non-suitable emotional feelings and to perform very simply document transferring as well as document archiving.

5 CONCLUSION

Our system provides sentimental analysis to calculate sentimental values for each document in order to visualize what kinds of emotional feelings are included in such a document. And it also performs to transfer the relevant documents without non-suitable emotional expressions as well as to archive such documents as PDF files, for example to form them through external LaTeXfacility. With our system, users, namely writers of documents, can realize document management effectively and efficiently.

REFERENCES

[1] T. Kumamoto, et al., “Proposal of Impression Mining from News Articles”, Lecture Notes in Computer Science Volume 3681, pp 901–910, 2005.

[2] M. Kyokane, et al., “A Study of Visualization for Hidden Relation between Published Documents and Message from Twitter,” International Journal of E-Learning and Educational Technologies in the Digital Media Vol.2, No.1, pp.217– 224, 2016.

[3] S. Doi, et al., “Calculation and Comparison of Sentimental Values for Tweeting Message during Multiple Years,” Proc. of The Second International Conference on Electronics and Software Science(ICESS2016@ Takamatsu), pp. 165 – 169, Nov. 2016.

Exploring Review Spammers by Review Similarity: A Case of Fake Review in Taiwan

Min-Yuh Day[1], Chih-Chien Wang[2], Chien-Chang Chen[1], Shao-Chieh Yang[2]
[1] *Tamkang University, Taipei, Taiwan*
[2] *National Taipei University, Taipei, Taiwan*
Email: myday@mail.tku.edu.tw, wangson@mail.ntpu.edu.tw, ccchen34@mail.tku.edu.tw, steven821015@gmail.com

Abstract

Understanding the phenomenon of spam reviews in social media is now an emerging and important issue since some enterprises may hire spammers to post fake reviews to promote their product or demote product of their competitors. The hired spammers are paid based on the fake reviews. Thus, these spammers may rewrite previous reviews as new review to earn the money. Thus, review similarity maybe a cue to detect fake reviews. Although literature had investigated the spam reviews, the review similarity of real review spammers is relatively unexplored. The objective of this paper is to explore the review spammers with a real case of fake reviews in Taiwan by investigating the cosine similarity and content length of reviews. We have proposed a text mining approach for a better understanding the phenomenon of fake reviews. The empirical results suggested that when comparing with normal reviews, the spam reviews were longer and with higher content similarity.

Keywords: Spam, Fake reviews, Text mining, Cosine similarity

1. Introduction

With the popularity of the Internet, Internet has penetrated into our daily life. People now express their comments and share their ideas about product, service or others on the Internet. These shared comments also serve as an important reference for other people to make decisions.

To encourage people to share their idea, most social media or product review website allow people to share/publish their comments or experience anonymously. The anonymous feature makes online space a place for "free-of-speech", in which people can say almost anything. Almost no one will screen the contents if the contents do not break the law and regulation.

Since many consumers will consult opinions from internet and people can provide their opinions without limitation, some unethic companies perceive the opportunistic opportunity of manipulating the majority of online opinion by providing online opinions. Thus, these unethical companies begin to hire spammers to publish fake reviews to promote reputation of themselves or to attack their competitors.

Because of the lack of appropriate filtering mechanisms on the Internet, fake reviews are flooding the Internet. In 2015, Amazon filed legal action against 1,114 spammers because of the fake reviews, which had seriously affected Amazon's goodwill, misled the consumer, and influenced the seller's trust to Amazon.

Spammers usually post in a short time with the intention to lead the opinions. They tend tomislead users to make inappropriate decision[1]. The spammers are usually part time or full time workers hired by companies to distribute fake reviews. They create fake reviews in exchange of payment. To save time and efforts, spammers may duplicate and modify previous reviews as new reviews. Thus, spam reviews may be similar with each others.

This study used a real case of fake reviews in Taiwan to discuss the similarity of fake reviews. We proposed the idea that the similarity could be a cue to detect fake reviews since that similarity among fake reviews are usually higher than that among ordinary reviews and similarity between fake reviews and ordinary reviews.

2. Related Works

2.1. Spam reviews

Spam refers to send bulk messages that the audiences do not want. In the age of internet, people get messages from e-mail, instant messaging, blog, news media, social networking, web search... and so on. They receive ordinary messages, advertising as well as spams from these media.

The history of spam can be traced back to the 1970s[2]. The initial idea of spams limits to spamming e-mail. However, due to the development of internet applications, there are new issue of spamming, such as spam in web search engine and social media.

Due to the rapid evolution of social media, social spam is now a great challenge. Chakraborty, et al. [3] argued that there were four kind of social spams. First, Malicious Links, which usually contains damage or fraud

link or other means to harm users or computers. Second, Fake Profiles, which usually provide fake personal information to avoid being found and tempted to keep in touch with the normal users. Third, Bulk Submissions, which contains a group of comments published multiple times with the same or similar text. Fourthly, Fraudulent Reviews, which claim that the product is good, or give a negative comment to attack products of competitors, even if the commenters did not have consumption experience on using the product.

Both bulk submissions and fraudulent reviews mentioned by Chakraborty, et al. [3] are relative with spam review. Since online reviews are a valuable message source for consumers to make purchase decision, companies are highly concerned with the online opinions to the product. However, not all online reviews are truthful and trustworthy since some of them are fake reviews by spammers. Previous studies had conduct review spam detection using various machine learning techniques [4].

Supervised learning were usually used to anti-fake review detection. Previous literature usually use review text itself and reviewer information as cue to detect fake review [4].

2.2. Text Mining and Similarity Analysis

Dang and Ahmad [5] mentioned that about 90% of real world data is unstructured. It is impractical to manually analyze the large number of unstructured textual information. Thus, as a result, text mining techniques are being developed to mechanize the process of analyzing this information.

Losiewicz, et al. [6] revealed that text mining architecture composed of three functions: Data collection, data warehousing, and data Exploitation. Each of these three functions included two sub-functions: Data collection contains data source selection and file selection; Data warehousing contains data conversion and data storage; Data exploitation contains data mining and data presentation. Similarity analysis in text mining are used to explore the degree of association between two documents or two sentences. We assume that the same person will have similar terms and characteristics in writing. Thus, we convert words in a posted review into tokens to calculate similarity of different posts. If the similarity of the two documents is higher, it means that there are more words in common in these two documents.

There are some approaches to calculate the similarity. For example, cosine similarity calculates the angle of the two vectors in the high dimension space; Jaccard similarity calculates the degree of similarity between the two sets; Euclidean distance calculates the actual distance between two vectors; Manhattan distance calculates the sum of the absolute wheelbase on the street map.

Lau, et al. [7] calculates Amazon's review similarity analysis using cosine similarity. In their study, if similarity above some threshold, they manually reviewed them to determine if they were spam or not.

Table 1. Fake and Normal Reviews in the Current Study

	Fake Reviews	Normal Reviews	Total
Post	434	7457	7891

Jindal and Liu [8] divides the word of mouth into three categories: Fake opinion, ordinary review, and non-comment. They collected 5.8 million reviews of products on Amazon generated by 2.14 million users and counted similarity by Jaccard similarity to judgment real or fake reviews. They got accuracy rate of 78% in their study.

Lin, et al. [9] collected 2000 fake reviews among 155080 normal reviews by Jaccard similarity. They regarded reviews as fake reviews if similarity was higher than or equal 0.7. They using those data to training the model by Logistic regression and SVM for fake review detection and try to use this model detecting other database. They got precision rate of 85% in their study.

Algur, et al. [10] used cosine similarity to detect fake from normal reviews. They considered duplicated reviews and near duplicated reviews as spam reviews, and regarded unique reviews as non-spam reviews.

As mention above, previous studies had used content similarity as feature for detecting fake reviews. However, previous studies only assume that duplicated and near duplicated reviews were fake reviews. Few previous studies, if any, had used real fake review case to compare the similarity among fake reviews and among ordinary reviews. Thus, this study uses a real case of fake reviews to reveal the correctiveness of previous assumption that some fake reviews are duplicated or near duplicated from previous reviews.

3.Methodology

3.1. Data Corpus

The data we used in this study were the same as that used in prior research [11, 12]. In this study, we focus on the original posts for understanding the fake review posts and normal posts. As shown in Table 1, we collected 7891 posts, including 434 fake review posts and 7457 normal posts

3.2. Analysis Methods

Figure 1 reveals a two phases analysis framework for understanding the phenomenon of spam review. Firstly, we calculate the number of Chinese characters in phase one. We calculate the average length of all posts, spam posts and normal posts. Then, content similarity scores were calculated to understand if the fake reviews are duplicated or near duplicated from other fake reviews.

ISBN: 978-1-941968-42-0 ©2017 SDIWC

Jieba (https://github.com/fxsjy/jieba), a Chinese segmentation tool, was used to segment Chinese contents into word tokens. We calculate similarity of words token by Cosine similarity.

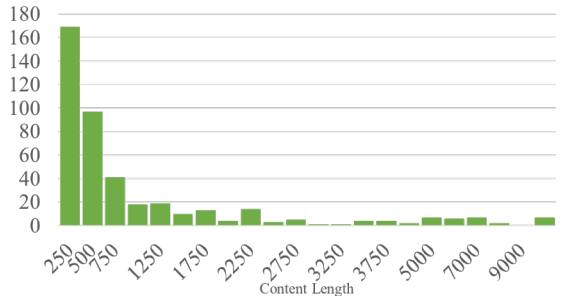
4. Data Analysis and Results

4.1. Content Length

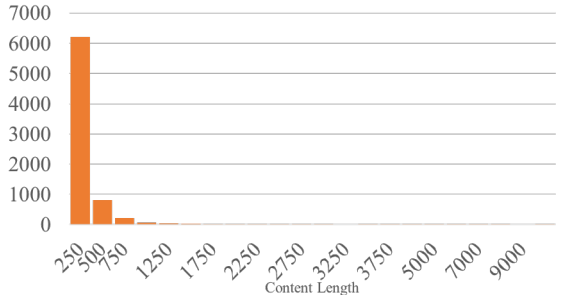
Table 2 shows the statistical summary of the content length of reviews. The mean length for all reviews posts is 236.93 Chinese characters. The mean length of spam reviews posts is 1058.91 Chinese characters, while the mean length (189.09) of normal review posts is 189.09 Chinese characters. In average, fake reviews were longer than the normal reviews.

Figure 2(a) shows the length distribution of spam reviews. In the collected 434 spam reviews, we found that 69 spam reviews are with less than 250 Chinese characters (38.94%), 265 spam reviews are with more than 250 Chinese characters (61.06%).

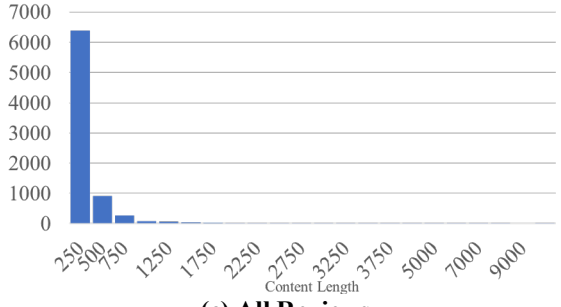
Figure 2(b) shows the length distribution of normal reviews. In the collected 7,457 normal reviews, we found that 6,207(83.24%) normal reviews were shorter than



(a) Spam Reviews



(b) Normal Reviews



(c) All Reviews

Figure 2 Length Distribution of the Reviews

Table 2 Statistical Summary of the Content Length of Reviews

	All posts (n=7891)	Spam posts (n=434)	Normal posts (n= 7457)
Mean	236.93	1058.91	189.09
Standard Deviation	639.85	1906.06	425.01
First Quartile	65	168	63
Median	110	346.5	105
Third Quartile	202	1006.25	188

250 Chinese characters Only 1250 normal reviews were longer than 250 Chinese characters (16.76%).

In Figure 2(c), a total of 7891 reviews were collected we found that among the 7891 reviews, most reviews (80.80%) content length was shorter than 250 Chinese characters. The results suggest that the length of review content is short.

The empirical analysis results show that the length of spam review is longer than normal reviews' review content in average. We argued that spammers might use long detailed reviews to persuade normal users.

4.2. Review Similarity

We use Cosine similarity measurement to analyze the similarity among reviews. We divided reviews into three groups by length: Short reviews (review length shorter than 250 Chinese Characters), middle reviews (review length between 251 and 750 Chinese Characters), and long reviews (review length longer than 750 Chinese Characters).

There were 169 fake reviews that were shorter than

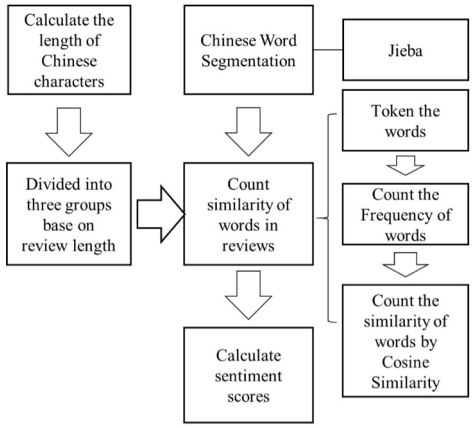


Figure 1 Content Similarity Analysis Procedure

250 Chinese characters (table 3). We randomly selected same amount (169) of normal posts for comparison purpose. In table 3, we found that the similarity among

Table 3 Similarity Analysis for Short Reviews (review length shorter than 250 Chinese Characters)

	(1) Similarity among Fake Reviews	(2) Similarity between Fake and Normal Reviews	(3) Similarity among Normal Reviews
Average Similarity	0.21	0.17	0.16
Standard Deviation	0.12	0.11	0.10
F value (P value)	F=664.09 (p<0.001)		
Post Hoc Test	(1)>(2)>(3)		

Notes: There were 169 short length fake reviews in our corpus. We randomly selected another 169 short length normal reviews for comparison purpose.

Table 4 Similarity Analysis for Middle Length Reviews (review length between 251 and 750 Chinese Characters)

	(1) Similarity among Fake Reviews	(2) Similarity between Fake and Normal Reviews	(3) Similarity among Normal Reviews
Average Similarity	0.37	0.32	0.30
Standard Deviation	0.15	0.15	0.15
F value (P value)	F=683.27 (p<0.001)		
Post Hoc Test	(1)>(2)>(3)		

Notes: There were 138 middle length fake reviews in our corpus. We randomly selected another 138 middle length normal reviews for comparison purpose.

Table 5 Similarity Analysis for Long Length Reviews (review length longer than 750 Chinese Characters)

	(1) Similarity among Fake Reviews	(2) Similarity between Fake and Normal Reviews	(3) Similarity among Normal Reviews
Average Similarity	0.50	0.39	0.34
Standard Deviation	0.23	0.23	0.21
F value (P value)	F=1060.00 (p<0.001)		
Post Hoc Test	(1)>(2)>(3)		

Notes: There were 127 long length fake reviews in our corpus. We randomly selected another 127 long length normal reviews for comparison purpose.

fake reviews were the highest, and the similarity among normal reviews are the lowest. The average similarity between fake review and normal reviews are in the middle. There were 138 fake reviews with content length in the range of 251 to 750 Chinese characters (Table 4). We randomly selected the same amount (138) of normal reviews for comparison purpose. In table 4, we also found that the similarity among fake reviews are the highest, and the similarity among normal reviews are the lowest. The average similarity between fake review and normal reviews are in the middle.

There were 127 fake reviews with content length of more than 751 Chinese characters (Table 5). We randomly selected the same amount (127) of normal reviews for comparison purpose. In Table 5, we found that the similarity among fake reviews are the highest, and the similarity of the group among normal reviews are the lowest. The average similarity between fake review and normal reviews are in the middle.

Table 5 reveals that the average similarity coefficient among long fake review is 0.50 while the average similarity coefficient is 0.34 among long normal review. The difference of similarity coefficient between long fake reviews and long normal review is 0.16. Table 6 reveals the similarity analysis results for all reviews (do not divide reviews into three groups). The average similarity coefficient among fake review is 0.33 (do not divide reviews into three groups), while the average similarity coefficient among long normal review is 0.25(do not divide reviews into three groups). Thus, review length is a potential moderator when using content similarity as a cue to detect fake reviews.

5. Discussion

Based on the content similarity analysis for fake review, we found that the similarity among fake reviews were higher than that among normal reviews or between normal reviews and fake reviews, no matter the review content is with short or long. However, if we did not divide the reviews based on length of the reviews, we can not observe this phenomenon of similarity since the similarity are low between long and short length reviews.

Secondly, we found that normal reviews with 250 or less Chinese characters accounted for 83.24% of normal reviews. However, spam posts with 250 or less Chinese characters accounted for only 38.94%. The research results suggest that the spam posts are generally more longer than normal posts.

The contributions of this paper are three folds. First, we have proposed a texting mining approach and explored the review spammers with a real case of fake review in Taiwan by investigating the cosine similarity and content length of reviews. We discovered the spam reviews tend to have higher content similarity and longer reviews than normal reviews

Second, we used text mining techniques with Cosine similarity for analyzing the similarity of spam reviews post. We found the content similarity among fake

Table 6 Similarity Analysis for All Reviews

	(1) Similarity among Fake Reviews	(2) Similarity between Fake and Normal Reviews	(3) Similarity among Normal Reviews
Average Similarity	0.33	0.27	0.25
Standard Deviation	0.20	0.18	0.17
F value (P value)	F=1685.86 (p<0.001)		
Post Hoc Test	(1)>(2)>(3)		

Notes: There were 434 fake reviews in our corpus. We randomly selected another 434 long length normal reviews for comparison purpose.

reviews are higher than the similarity between fake and normal reviews and similarity among normal reviews.

Third, based on the analysis of content length, spam reviews are longer than normal reviews. This empirical analysis results suggest the fact that spammers would likely to use longer and detailed contents to persuade the consumers to believe the review content they post on Internet. Based on this observation, we should keep more attention on long length reviews when we want to detect fake reviews.

6. Reference

[1] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review spam detection via temporal pattern discovery," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 823-831: ACM.

[2] F. Brunton, *Spam: A shadow history of the Internet*. MIT Press, 2013.

[3] M. Chakraborty, S. Pal, R. Pramanik, and C. R. Chowdary, "Recent developments in social spam detection and combating techniques: A survey," *Information Processing & Management,* vol. 52, no. 6, pp. 1053-1073, 2016.

[4] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," *Journal of Big Data,* vol. 2, no. 1, p. 23, 2015.

[5] D. S. Dang and P. H. Ahmad, "A Review of Text Mining Techniques Associated with Various Application Areas," *International Journal of Science and Research (IJSR),* vol. 4, no. 2, pp. 2461-2466, 2015.

[6] P. Losiewicz, D. W. Oard, and R. N. Kostoff, "Textual data mining to support science and technology management," *Journal of Intelligent Information Systems,* vol. 15, no. 2, pp. 99-119, 2000.

[7] R. Y. Lau, S. Liao, R. C.-W. Kwok, K. Xu, Y. Xia, and Y. Li, "Text mining and probabilistic language modeling for online review spam detection," *ACM Transactions on Management Information Systems (TMIS),* vol. 2, no. 4, p. 25, 2011.

[8] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 2008, pp. 219-230: ACM.

[9] Y. Lin, T. Zhu, H. Wu, J. Zhang, X. Wang, and A. Zhou, "Towards online anti-opinion spam: Spotting fake reviews from the review sequence," in *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, 2014, pp. 261-264: IEEE.

[10] S. P. Algur, A. P. Patil, P. Hiremath, and S. Shivashankar, "Conceptual level similarity measure based review spam detection," in *Signal and Image Processing (ICSIP), 2010 International Conference on*, 2010, pp. 416-423: IEEE.

[11] C.-C. Wang, M.-Y. Day, and Y.-R. Lin, "Toward understanding the cliques of opinion spammers with social network analysis," in *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, 2016, pp. 1163-1169: IEEE.

[12] C.-C. Wang, M.-Y. Day, and Y.-R. Lin, "A Real Case Analytics on Social Network of Opinion Spammers," in *Information Reuse and Integration (IRI), 2016 IEEE 17th International Conference on*, 2016, pp. 623-630: IEEE.

Tobacco Leaf Area Growth Simulation with the Variational Level Set Method

Israel Pineda and Oubong Gwun
Chonbuk National University
Jeonju - South Korea
igpaec@gmail.com, obgwun@jbnu.ac.kr

***Abstract*—In this paper, we propose a simulation of the area growth process of the tobacco leaf. A signed distance function represents the leaf, and the Variational Level Set Method tracks the evolution of the interface between the leaf and the background. The evolution of that interface is guided by an energy function that depends on internal and external characteristics. The internal energy helps to keep the interface as a signed distance function that avoids reinitialization and therefore enhances the result. The external energy helps to recreate the shape of the tobacco leaf using a reference image as input. We also discuss a special treatment for the senescence stage of the leaf and describe the method and its configuration parameters. In this paper, the focus is on replicating the area growth pattern. Therefore, we measure the area and compare it with real data from plant measurements to validate our work. We use L_1 error to numerically assess the output.**

***Index Terms*—Tobacco leaf simulation, level set method, leaf growth**

I. Introduction

Simulation of plants and their different parts provides useful insights to scientists and engineers across different disciplines. Biology applications, for example, use this kind of simulation to create new data, validate data from measurements, and validate experimental procedures. Agriculture applications might use simulation for production planning and cost estimation. In computer science, especially in computer graphics applications, the simulation of leaves can be used to create animations for movie production, games, virtual worlds, and augmented reality applications. In this paper, we concentrate on replicating the pattern of the area growth during the development of the tobacco leaf, so only the visual plausibility of the leaf simulation will be evaluated. Despite all the possible applications, describing this process with a unified model is difficult because the growth process of a leaf involves sophisticated chemical and biological interactions at different scales.

A particularly interesting case is the tobacco leaf. Tobacco is a plant with economic, health, and social repercussions. It is widely consumed around the world mainly through derivatives of its leaves. Government and health organizations are constantly conducting studies related to tobacco production and consumption. New tools to understand the nature of this leaf might help to alleviate the problems it generates. The World Health Organization named tobacco as the leading preventable cause of death in the world [6]. We believe that a simulation of the growth of the tobacco leaf might be a useful tool to such organizations and to scientists and engineers who study the different aspects of the leaf.

In this paper, we present a simulation that replicates the area growth of the tobacco leaf. We propose a representation method for the leaf as well as the necessary numerical mechanism to evolve the given representation over time. The proposed simulation provides a visually plausible reproduction of the growth of the area of the leaf. Other measurements are not considered in this study.

The proposed method uses an implicit distance function to represent the shape of the leaf; only its boundary is considered. Then, the Variational Level Set Method formulation is used to evolve the implicit function over time until it matches the shape of the leaf in a given input image. We discuss the details of the method and its origins in Section III.

We present and discuss the results of several experiments that were executed in Section IV. We also compare our results with real data from the leaf measurements by Suggs [11]. At the end of this paper, we provide conclusions and some ideas that we would like to address in the future.

II. Related Work

The main method to simulate the growth of plants and leaves, for computer graphics applications, is arguably the use of L-Systems [8], they are also one of the oldest methods available. L-Systems are rule-based programs that can replicate important growth processes based on the principle of self-similarity. They work by executing procedures that create the next state of the object, usually a plant or a leaf. However, coming up with ad hoc rules for every kind of leaf is difficult.

Another method is to use triangle meshes to represent the leaf. The vertices of the triangles should evolve according to some external logic. Alsweis et al. [1] conducted a study using this method. They represented the geometry with a triangle mesh, and then an incompressible fluid simulation tracked the development of the leaf.

Most of the techniques to simulate leaves use a triangle mesh at some point to represent the necessary geometry; this kind of representation is called explicit. The difficulty with this kind of representation is that all the calculations should be performed at every point of the mesh to obtain the output from the simulation. This is computationally expensive. Another limitation is the lack of ability to deal with topological changes of the objects during the simulation.

ISBN: 978-1-941968-42-0

A different approach is the use of an implicit representation. This means that instead of using a mesh to represent the geometry, an implicit function, specifically an implicit distance function, can be used. This kind of function provides the distance from every point in the computational domain to the nearest point on the interface of the object. A sign can also be assigned to every point; this is used to quickly identify between the inside and the outside of the interface. The common convention is to have a positive value outside and a negative value inside the object. When the sign is used, this representation is called a signed distance function. Detailed information on signed distance functions can be found on the work by Jones et al. [3].

The signed distance function, as described before, represents the leaf that needs to be simulated. A method that allows for the evolution of the leaf over time is then necessary. That means that the interface will move. This is achieved by the Level Set Method [7], [10], which uses the solution of partial differential equations to calculate the new state of the interface. The Level Set Method should use a well-defined signed distance function. Otherwise, undesirable effects might occur.

One interesting case of the Level Set Method is the so-called Variational Level Set Method similar to the one presented by Li et at. [4] that avoids the use of reinitialization. The idea of the Variational Level Set Method is that it uses calculus of variations to define the evolution equation. The work of Li et at. [4] builds on top of the method proposed by Chan and Vese [2].

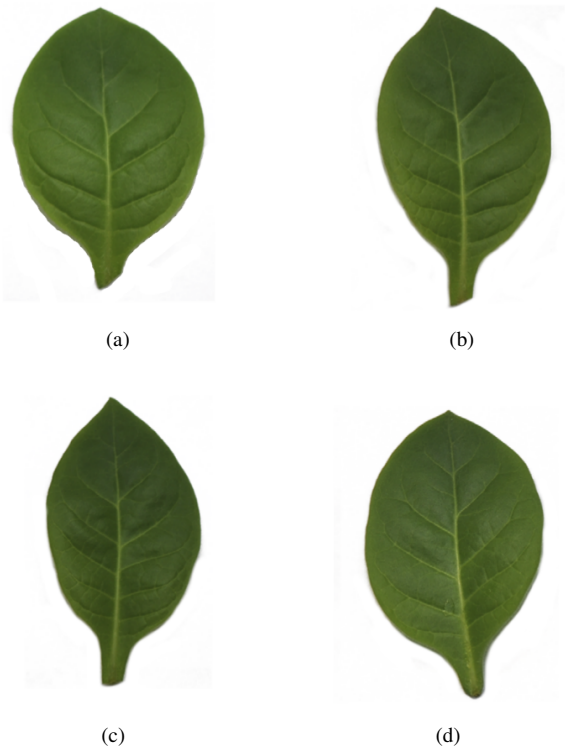


Figure 1: Different exemplars of tobacco leaves.

III. Method

In this section, we describe a method to simulate the growth of the leaf. After defining some prerequisites, we provide the necessary details and explain possible configurations. We also explain a variation of the method that allows for the simulation of the senescence of the leaf.

A. Prerequisites

The method requires an input image that contains a picture of a mature tobacco leaf. We call this image I and it is used as a reference to create the desired shape. An important assumption is made at this point. We assume that the image only contains a well-identifiable picture of the target leaf. We make this assumption because we want to avoid the hurdles of segmentation, which is not the goal of this paper.

Representing image I using the $YCbCr$ color space is important. This allows for an easy extraction of the grayscale image, which in this case is simply the Y component. If another color space is used, RGB for example, the respective transformation from color to grayscale should be performed. We will later use image I to guide the evolution.

B. Leaf Growth

In this simulation, a signed distance function represents the boundary of the target leaf. This function provides the distance from any point in the computation domain to the nearest point on the leaf. The sign of this function also indicates if a given point is inside or outside the leaf. We use the normal convention: positive outside and negative inside.

An initial distance function is needed to start with. We proposed to use a geometrically generated object to initialize the distance function. A circle is a simple way to create the initial function because its analytic representation is well-known. We call this initial interface ϕ_0. We discuss other initialization possibilities in Subsection III-C.

The initial interface ϕ_0 needs to move toward the leaf depicted in the reference image I. We discuss how this evolution is enforced later in this section. However, it is possible to see that at some moment the evolution should stop. That should happen when the interface reaches the edge of the leaf. For this reason, we need to define an edge detector. The edge detector is defined as follows:

$$g = \frac{1}{1 + |\nabla G_\sigma * I|^2}, \tag{1}$$

where a Gaussian convolution is applied to the image I and then the gradient of the image is taken in the denominator. The edge detector becomes zero when it reaches the edge of the leaf and a value bigger than zero everywhere else. We use this value later on.

At this point, the signed distance function ϕ represents the interface of the leaf, and the edge detector g indicates when the simulation should stop. With this information, we can work on the evolution of the interface ϕ over time.

ISBN: 978-1-941968-42-0 ©2017 SDIWC

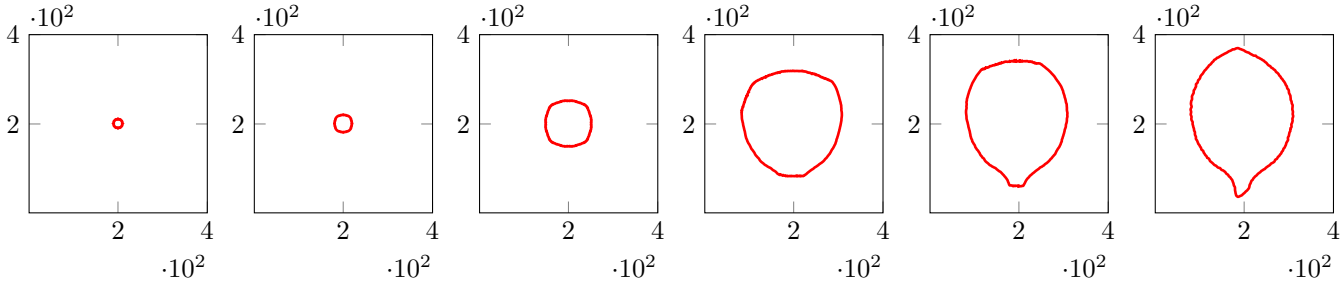


Figure 2: Interface evolution with respect to the input reference image at different times during the simulation.

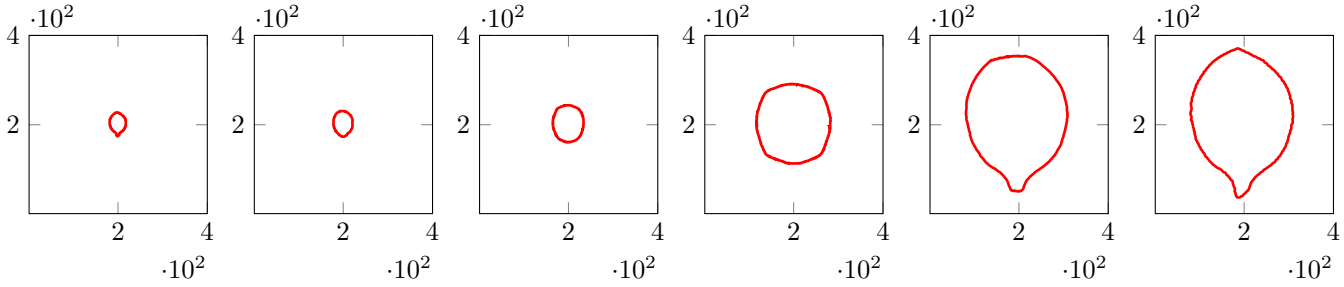


Figure 3: Usage of small leaf for the initialization of the distance function and senescence.

Here, we describe how the interface moves from ϕ^n to ϕ^{n+1}. We follow the ideas from Li et al. [4] in this section. We start by defining the following total energy functional:

$$\mathcal{E}(\phi) = \mu\mathcal{P}(\phi) + \mathcal{E}_{g,\lambda,\nu}(\phi). \tag{2}$$

Eq. (2) contains two terms, each representing the internal and external energy. We start by defining the internal energy as follows:

$$\mathcal{P}(\phi) = \int_\Omega \frac{1}{2}(|\nabla\phi| - 1)dxdy. \tag{3}$$

The external energy will then use the information from the reference image I using the edge detector g. The external energy is defined as follows:

$$\mathcal{E}_{g,\lambda,\nu}(\phi) = \lambda\mathcal{L}_g(\phi) + \nu\mathcal{A}_g(\phi). \tag{4}$$

This external energy also contains two components. Each one is controlled by a regularization term λ and ν. The first term deals with the length of the interface. This term is defined as

$$\mathcal{L}_g(\phi) = \int_\Omega g\delta(\phi)|\nabla\phi|dxdy, \tag{5}$$

where δ is the Dirac delta function of ϕ.

The second term of the external energy function depicted in Eq. (4) deals with the area covered by the current state of the interface. It is defined as

$$\mathcal{A}_g(\phi) = \int_\Omega gH(-\phi)dxdy, \tag{6}$$

where H is the Heaviside function of ϕ.

After fully defining the functional, the evolution equation can be provided.The evolution equation is a variation from the family of Hamilton-Jacobi equations and is defined as follows:

$$\frac{\partial\phi}{\partial t} + \frac{\partial\mathcal{E}}{\partial\phi} = 0 \tag{7}$$

Discretization techniques for this equation along with additional information can be found in the work of Li et al. [4].

We mention now a special consideration. In the work of Li et al. [4], only the final step is relevant. However, in our case, every intermediate state of ϕ is important. For this reason, we have to set the time steps of the simulation to be sufficiently small to produce a smooth transition between different times; even if the gradient is low, we still keep a conservative time step. We enforce this condition using a Courant-Friedreichs-Lewy (CFL) condition with a CFL number of 0.5.

Following the ideas discussed so far, we have a function ϕ that grows from an initial state toward the information provided by the reference image I. We track all the intermediate states to create a sequence of the growth process. We also keep track of the area growth to later validate our results. The growth process is determined by Eq. (7).

The results of experiments with the proposed method are shown in the next section.

C. Senescence

The behavior of the leaf during the last stage of its life is rather different from the development stage of the leaf. Fig. 4 (a) shows real data from the development of the area of the tobacco leaf. After the leaf has reached its maximum area toward the end, fluctuation in the area occurs. This is called the senescence of the leaf. To simulate the behavior during the senescence stage, we work with the ν parameter described before. The change of the sign of this value allows for the evolution of the interface to go in the opposite direction, thereby reducing the area of the leaf.

We use a configuration array with a set of switching times; we call this array S. Every time we reach a time specified in

the array, we change the sign of ν according to the following logic:

$$\forall t \in S : \nu = \nu * (-1). \tag{8}$$

For more accuracy and realism, S is filled with information based on observations from the data presented in Fig. 4 (a).

There is a caveat to the inclusion of the senescence logic into the simulation. If the final area of the leaf has been already reached and the interface is located exactly at the interface of the leaf, the change of sign has no effect. Therefore, this switch should happen before the area reaches its maximum.

IV. Results

In this section, we present the results of our experiments with the proposed method. We show the evolution of the interface visually and numerically. We also compare these results with real measurements of the area growth of the tobacco leaf.

In all the experiments, we used tobacco leaves. Fig. 1 shows the input images to be used as the reference image I. These images were obtained from the work of Lu et at. [5]. The experiments used images that contain only the leaf without any other object, shadow, or background.

The configuration of the simulation is as follows: The initial interface is defined by a circle. The output depends on the location of the center of the circle and its radius. The center of the initial circle was at $x = 200$, $y = 200$, which coincides with the center of the reference image I. The radius was 10 pixels. All the experiments used reference images with a resolution of 400×400 pixels with sound results. This is a rather low resolution by current standards; however, our method works effectively under this condition. The parameters for the energy function were $\mu = 0.04$, $\nu = -1.5$, $\lambda = 5$, $\sigma = 1.5$, and $\epsilon = 1.5$.

With this configuration, we show how the interface evolves over time toward the leaf. Fig. 2 shows the growth process from the initial circle until the interface matches the shape of the leaf depicted in the reference image. It also shows several intermediate states. Using the previously mentioned parameters, the simulation required 800 iterations to reach the final state. Depending on the location of the interface and the selected parameters, the simulation might required more or fewer iterations to complete its work. For example, the increase in the absolute value of the parameters ν and λ might speed up the process. However, these changes also produce undesired effects. For instance, an exaggerated increase of ν and λ produces an abrupt movement of the interface that makes the simulation miss the boundary and keep evolving indefinitely. Therefore, care must be taken in the selection of the parameters.

Fig. 3 shows a different experiment. Similar parameters were used for the configuration. However, initialization is different. In this case, ϕ_0 was configured using a signed distance function extracted from a small leaf in its early stage. The goal of such initialization is to have a better approximation of the real growth process. However, our experiments show that there is little difference because the interface evolves

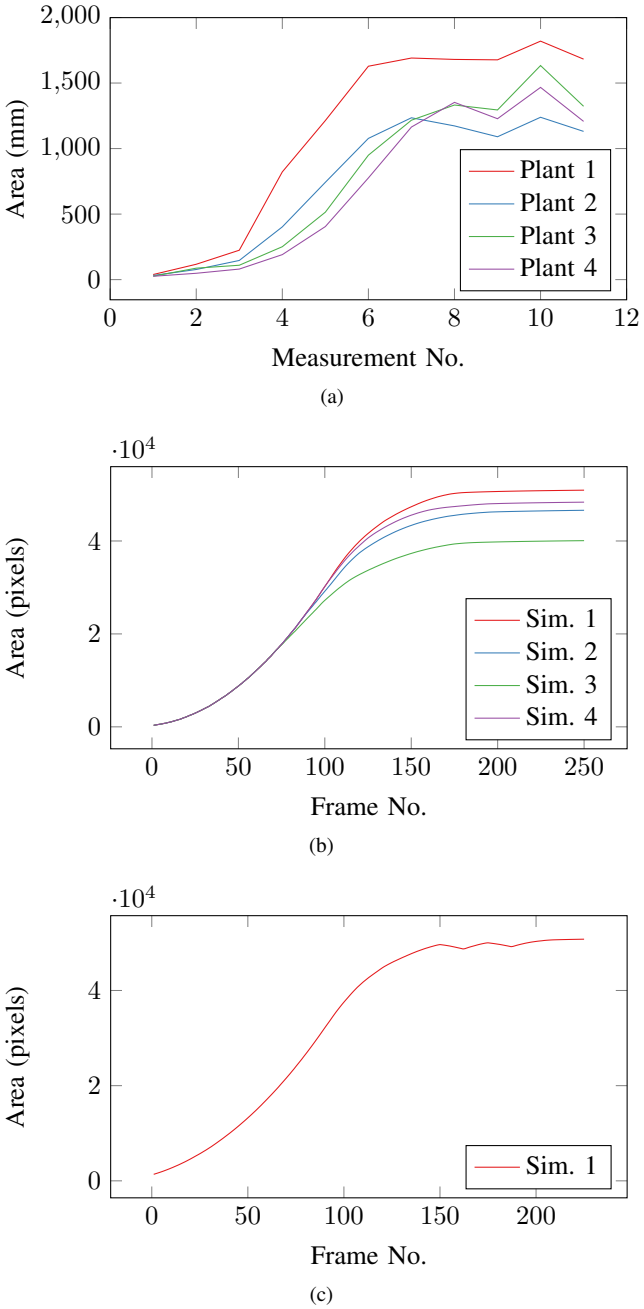


Figure 4: (a) Real data. (b) Simulated data. (c) Senescence simulation data.

toward the normal direction and the shape of the small leaf is lost after some execution time. Thus, even though some small difference occurs in the beginning, the results tend to be similar toward the end of the process.

Fig. 4 (a) shows real data corresponding to the area growth of the tobacco leaf from the measurements carried out by Suggs [11]. Fig. 4 (b) shows the data obtained from our experiments executing the proposed method. This S-shape pattern is typical of the growth of plants and is usually formalized using the logistic function or its variation the generalized logistic function [9]. The domain and the range of both figures are different because our simulation is working

with pixel sizes instead of the real size of the leaf; in this work, we deal only with replicating the pattern of the area growth. Fig. 4 (c) shows the results of the simulation when the senescence logic is included, as described in Subsection III-C. Fig. 4 (c) shows additional details toward the end of the simulation when the senescence stage is also considered.

We calculated the accuracy of the results using first-order L_1 error. The simulated data was downsampled to have the same number of elements as the number of real measurements. Then, we changed the range of the simulated data to match the range of the measurements. The L_1 error was calculated as

$$\frac{1}{A_{max}} \int |A_m - A_s|, \tag{9}$$

where A_m is the data from measurements and A_s is the data generated from the simulation. We used the maximum area A_{max} of the real data to normalize the results. Using this error measurement for plant number one, the simulation yielded an error value of $L_1 = 0.524$ without senescence and $L_1 = 0.551$ with senescence.

Careful selection of parameters can avoid undesired effects. For example, we found that when the interface touches the boundary too soon, the logistic growth is lost and the growth behaves linearly.

After obtaining the results shown so far, additional techniques could be applied to enhance the visual appeal of the results. The implicit representation of the leaf is well suited for rendering techniques such as raytracing. However, the rendering of these objects is out of the scope of this study, and we mention it just for the sake of completeness.

V. Conclusions

We proposed a framework based on the idea of the Variational Level Set Method to create a simulation of the area growth of the tobacco leaf. The results showed the shape of the constructed interface at different times during the simulation as well as the numerical tracking of the area over time. The experiments used two initialization strategies: geometrical-based initialization and initialization based on a small leaf. Additionally, the proposed method included a special logic to replicate the behavior of the leaf during the senescence stage. This was possible to achieve by the proper manipulation of the Variational Level Set Parameters in combination with a set of switching times. The results showed that the proposed method can replicate the pattern of the growth process.

Other species of plants might also benefit from this kind of technique. However, ad hoc modifications might be necessary. In our case, we restricted the experiments to the tobacco leaf because of the availability of real data to validate the simulation.

In this study, we followed the ideas from Li et al. [4] where the Variational Level Set Method was formulated. Chan and Vese [2] presented a similar approach. Both of these techniques were developed for image segmentation. In our work, we have adopted these ideas to replicate the growth process of the tobacco leaf, which opens the door to a new way to create animations of plants as well as new engineering applications.

We briefly mentioned some post-processing techniques that might help to obtain a more appealing visualization of the output. For example, the reference image can be used to create a texture map, and then the leaf could be rendered to obtain the final result.

We would like to explore several directions in future works. For example, we would like to improve the accuracy of the shape during the evolution. This might need the modification of the energy function with botanical information. The internal energy function would especially provide a conducive place to include additional information from the target leaf. In general, the creation of more sophisticated energy functions is particularly interesting. Another interesting case of study is the drying process of the tobacco leaf, we would also like to investigate in this direction.

References

[1] Monssef Alsweis and Oliver Deussen. Procedural techniques for simulating the growth of plant leaves and adapting venation patterns. In *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology - VRST '15*, pages 95–101, New York, New York, USA, 2015. ACM Press.

[2] Tony Chan and Luminita Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001.

[3] Mark Jones, Andreas Bærentzen, and Milos Sramek. 3D distance fields: a survey of techniques and applications. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):581–599, jul 2006.

[4] Chunming Li, Chenyang Xu, Changfeng Gui, and Martin Fox. Level Set Evolution without Re-Initialization: A New Variational Formulation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 430–436. IEEE, 2005.

[5] Jie Lu, Zhi-Xin Du, Jun Kong, Ling-Na Chen, Yan-Hong Qiu, Gui-Fen Li, Xiao-Hua Meng, and Shui-Fang Zhu. Transcriptome Analysis of Nicotiana tabacum Infected by Cucumber mosaic virus during Systemic Symptom Development. *PLoS ONE*, 7(8):e43447, aug 2012.

[6] World Health Organization. WHO Report on the Global Tobacco Epidemic 2008 The MPOWER package. Technical report, Geneva, 2008.

[7] Stanley Osher and Ronald Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*. 2003.

[8] Przemyslaw Prusinkiewicz and Aristid Lindenmayer. *The algorithmic beauty of plants*. Springer-Verlag, 1990.

[9] F. J. Richards. A Flexible Growth Function for Empirical Use. *Journal of Experimental Botany*, 10(2):290–301, 1959.

[10] James Sethian. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences of the United States of America*, 93(4):1591–1595, feb 1996.

[11] Charles Suggs. Tobacco leaf area versus stalk diameter research. *Special Collections Research Center at NCSU Libraries*, 1961.