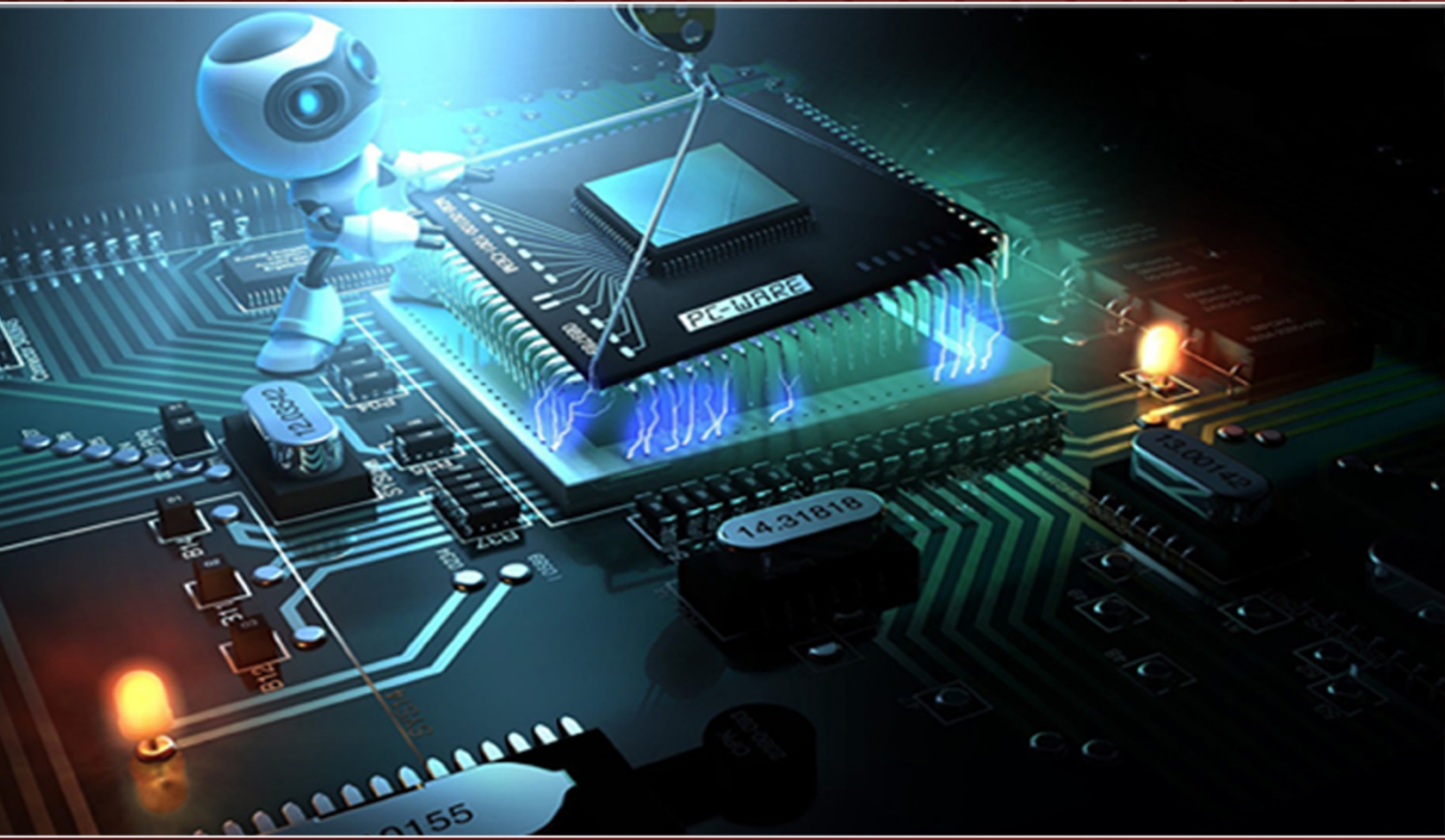
ISSN: 2054-7390

Volume 2. Issue 3 - June 2014

Transactions on Machine Learning and Artificial Intelligence



Society for Science and Education
UNITED KINGDOM

TABLE OF CONTENTS

EDITORIAL ADVISORY BOARD

DISCLAIMER

All the contributions are published in good faith and intentions to promote and encourage research activities around the globe. The contributions are property of their respective authors/owners and the journal is not responsible for any content that hurts someone's views or feelings etc.

TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 2, ISSUE 3
ISSN: 2054 - 7390

Image Learning of Electric Characteristics of Resistance, Capacitance, Inductance, and their Circuits by Java Programming

Masami Morooka[1], Suhua Qian[1], Midori Morooka[2]
[1] *Department of Electrical Engineering, Fukuoka Inst. Tech,higashi-ku, Fukuoka 811-0295, Japan;*
[2] *Flash Design Center, Micron Japan, Kamata 5-37-1, Ota-ku, Tokyo144-8721, Japan;*
morooka@ee.fit.ac.jp

ABSTRACT

Java programs with a graphical user interface (GUI) environment have been developed for an image learning of electric characteristics of resistance (R), capacitance (C), inductance (L), and their circuits. Text fields of selected parameters for the numerical calculation of the differential equations to describe the electric characteristics of the circuit are set on the display, such as the frequency of applied voltage and the values of R, C, L, time increments, and calculation times. The calculation used Runge-Kutta method is initiated by clicking the start button after inputting the desired values into the text fields. The calculated results are plotted immediately after the completion of the calculation as a figure on the display for the simulation of the electric characteristics of the circuit, such as changes in the voltages and currents with time. By changing the values in the text fields, the new results are represented immediately, and the new electric characteristics of the new circuit can be easily simulated. These Java programs are useful in education applications for rapidly and accurately image learning for the electric characteristics of the circuit. This program is also useful for learning the phenomena expressed by ordinary differential equations.

Keywords: Image learning, characteristics of electric circuit, resistance – capacitance – Inductance, Java programming, voltage – current characteristics of RCL circuits.Introductin

1 INTRODUCTION

In many situations, it is difficult to obtain the variation of the current and voltage in a circuit composed of a resistance (R), a capacitance (C), and an inductance (L). It is even more difficult to visualize the variation in the characteristics of the circuits when they are limited by each component. The electric characteristics of the circuit can be expressed by ordinary differential equations, which can be solved numerically using the Runge-Kutta method. With the wide use

DOI: 10.14738/tmlai.23.211
Publication Date: 9th June 2014
URL: http://dx.doi.org/10.14738/tmlai.23.211

of Java programming in a GUI (graphical user interface) environment and the rapid development of personal computers, the rapidly and accurately image learning for charge motions in electric and magnetic fields [1] and that for complicated diffusion of Au into Si [2] have been developed by the Java programming. In this paper, the ordinary differential equations for RCL circuits are solved numerically, and the voltage-current characteristics of the circuits are easily and accurately simulated using Java in a GUI environment. These Java programs are useful in education applications for rapidly and accurately image learning of electric characteristics of R, C, L, and their circuits.

2 NUMERICAL METHOD FOR ELECTRIC CIRCUITS

2.1 Basic Equations for Voltage-Current Characteristics of R, C, and L

The voltage $v(t)$ – current $i(t)$ characteristics of R, C, and L are given as

$$i_R(t) = \frac{v_R(t)}{R}, \tag{1}$$

$$i_C(t) = \frac{dq(t)}{dt} = C\frac{dv_C(t)}{dt}, \tag{2}$$

$$v_L(t) = L\frac{di_L(t)}{dt}, \tag{3}$$

where, t is time and $q(t)$ is the charge of capacitance, and the subscripts refer to each of the components.

2.2 Differential Equations and Numerical Method for Series RCL Circuits

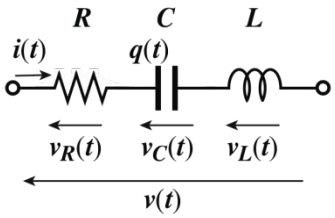


Figure 1: A series RCL circuit

The electric characteristics of a series circuit with an applied voltage $v(t)$ shown in Figure 1 are given by two ordinary differential equations:

$$\frac{dq(t)}{dt} = i(t), \tag{4}$$

$$\frac{di(t)}{dt} = \frac{v(t) - Ri(t) - q(t)/C}{L}. \tag{5}$$

In this case, the increment functions for the forth-order Runge-Kutta method are

$$k_1^{(1)} = i_j \,, \tag{6}$$

$$k_2^{(1)} = \frac{v(t) - Ri_j - q_j / C}{L} \,, \tag{7}$$

$$k_1^{(2)} = i_j + k_2^{(1)} h / 2 \,, \tag{8}$$

$$k_2^{(2)} = \frac{v(t + h/2) - R(i_j + k_2^{(1)} h/2) - (q_j + k_1^{(1)} h/2)/C}{L} \,, \tag{9}$$

$$k_1^{(3)} = i_j + k_2^{(2)} h / 2 \,, \tag{10}$$

$$k_2^{(3)} = \frac{v(t + h/2) - R(i_j + k_2^{(2)} h/2) - (q_j + k_1^{(2)} h/2)/C}{L} \,, \tag{11}$$

$$k_1^{(4)} = i_j + k_2^{(3)} h \,, \tag{12}$$

$$k_2^{(4)} = \frac{v(t + h) - R(i_j + k_2^{(3)} h) - (q_j + k_1^{(3)} h)/C}{L} \,, \tag{13}$$

where, h is the increment of t and the subscript $_j$ represents the known variables at a given t. The details on the calculation of the increment functions are described in the Reference [1]. The unknown variables q_{j+1}, and i_{j+1} at $t + h$ are given as

$$q_{j+1} = q_j + \frac{1}{6}(k_1^{(1)} + 2k_1^{(2)} + 2k_1^{(3)} + k_1^{(4)}) \,, \tag{14}$$

$$i_{j+1} = i_j + \frac{1}{6}(k_2^{(1)} + 2k_2^{(2)} + 2k_2^{(3)} + k_2^{(4)}) \,. \tag{15}$$

If the variables at a certain t are known, the numerical values at $t + h$ can be obtained from Equations (14) and (15), and then the values for $t + 2h$, and $t + 3h$ and so on, can be obtained by repeating the calculations. The voltage for each of the components can be obtained using the calculated q and i with

$$v_R(t) = Ri(t) \,, \tag{16}$$

$$v_C(t) = \frac{q(t)}{C} \,, \tag{17}$$

$$v_L(t) = v(t) - v_R(t) - v_C(t) \,. \tag{18}$$

2.3 Differential Equations and Numerical Method for Parallel RCL Circuits

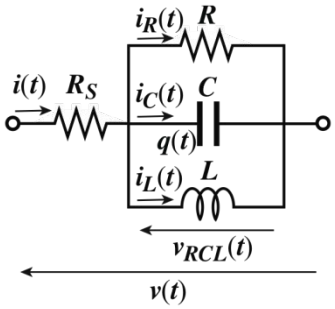


Figure 2: A parallel RCL circuit

The capacitance current depends on the change of the applied voltage, $dv(t)/dt = \Delta v/\Delta t$ as shown in Equation (2). If a voltage is applied directly to the capacitance, the capacitance current results in a very large numerical value depending on *h*. This occurs when the applied voltage is switched on or off, or a rectangle voltage is applied. Therefore, it is ideal to connect a series resistance that has a very small resistivity, in series with the capacitance for the numerical calculations, as shown R_S in Figure 2. The electric characteristics of the parallel circuit shown in Figure 2 are given by two ordinary differential equations using $v_{RCL}(t) = q(t)/C$:

$$\frac{di_L(t)}{dt} = \frac{q(t)}{CL}, \tag{19}$$

$$\frac{dq(t)}{dt} = \frac{v(t)}{R_S} - \frac{q(t)}{R_S C} - \frac{q(t)}{RC} - i_L(t). \tag{20}$$

In this case, the increment functions for the forth-order Runge-Kutta method are

$$k_1^{(1)} = \frac{q_j}{CL}, \tag{21}$$

$$k_2^{(1)} = \frac{v(t)}{R_S} - \frac{q_j}{R_S C} - \frac{q_j}{RC} - i_{Lj}, \tag{22}$$

$$k_1^{(2)} = \frac{q_j + k_2^{(1)} h/2}{CL}, \tag{23}$$

$$k_2^{(2)} = \frac{v(t+h/2)}{R_S} - \frac{q_j + k_2^{(1)} h/2}{R_S C} - \frac{q_j + k_2^{(1)} h/2}{RC} - (i_{Lj} + k_1^{(1)} h/2), \tag{24}$$

$$k_1^{(3)} = \frac{q_j + k_2^{(2)} h/2}{CL}, \tag{25}$$

$$k_2^{(3)} = \frac{v(t+h/2)}{R_S} - \frac{q_j + k_2^{(2)} h/2}{R_S C} - \frac{q_j + k_2^{(2)} h/2}{RC} - (i_{Lj} + k_1^{(2)} h/2), \tag{26}$$

$$k_1^{(4)} = \frac{q_j + k_2^{(3)} h}{CL}, \tag{27}$$

$$k_2^{(4)} = \frac{v(t+h)}{R_S} - \frac{q_j + k_2^{(3)} h}{R_S C} - \frac{q_j + k_2^{(3)} h}{RC} - (i_{Lj} + k_1^{(3)} h). \tag{28}$$

The unknown variables i_{Lj+1} and q_{j+1} at $t + h$ are given as

$$i_{Lj+1} = i_{Lj} + \frac{1}{6}(k_1^{(1)} + 2k_1^{(2)} + 2k_1^{(3)} + k_1^{(4)}), \tag{29}$$

$$q_{j+1} = q_j + \frac{1}{6}(k_2^{(1)} + 2k_2^{(2)} + 2k_2^{(3)} + k_2^{(4)}). \tag{30}$$

Other variables can be obtained using the calculated i_L and q:

$$v_{RCL}(t) = \frac{q(t)}{C}, \tag{31}$$

$$i(t) = \frac{v(t) - v_{RCL}(t)}{R_S}, \tag{32}$$

$$i_R(t) = \frac{v_{RCL}(t)}{R}, \tag{33}$$

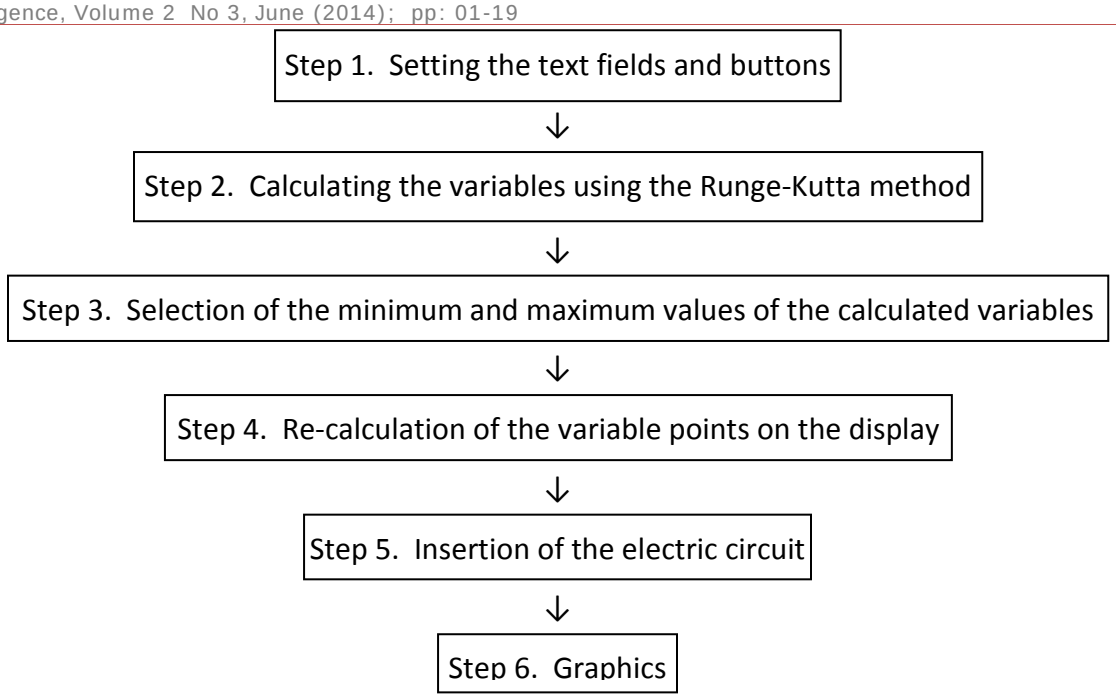
$$i_C(t) = i(t) - i_R(t) - i_L(t). \tag{34}$$

3 PROGRAMMING

In a numerical calculation, values that alternate with time, such as the applied voltage, can be conveniently expressed as cosine functions instead of sine functions. For example, $v(t) = v_0 \cos(2\pi ft)$. Here, f is the frequency of the applied voltage. If we use $v(t) = v_0 \sin(2\pi ft)$, then $v = 0$ at $f = 0$ despite the existence of a direct electric voltage $v = v_0$. A rectangle voltage, which corresponds to switch on and off the DC voltage, can be expressed using the cosine function as shown below.

```
t[0] = 0.0;
v[0] = 0.0;
for (i=0;i<=n-1;i++) {
        t[i+1] = h+t[i];
        fai = Math.cos(2.0*Math.PI*f*t[i+1]);
        if (fai>0) {
                fai=1.0;
        }
        if (fai<0) {
                fai=-1.0;
        }
        v[i+1] = v0*fai;
}
```

The main flow of the programming is shown below.



The basic programming for Steps 1 - 4 and Step 6 are same as the programming for the charge motion in electric and magnetic fields, and their details are previously reported [1]. In addition, it is useful to show the circuit as shown in Figures 1 and 2, while learning about the electric characteristics of RCL circuit (Step 5). A basic example of programming in Java to print L, R and C components and their values is shown below. Each component is prepared using only g.drawArc and g.drawLine statements. The functions of the main parts of the program are written as comments. The results from the program are shown in Figure 3.

```
/* RCL figure on display by java 2014.02.06 M. Morooka */
/*<applet code="FigRCL.class" width=700 height=600></applet>*/
import java.applet.Applet;
import java.awt.*;
public class FigRCL extends Applet {
        public void paint(Graphics g) {
                int yh;//height of figure
                yh=60;
//inductance
                int yL0,xLs,xLf,xw,rD,xLW,degree,xLl,xLr;
                yL0=yh/2+50;//y-center of L
                xLs=140;//left of L
                xw=yh*2/4;
                rD=xw*1/2+xw/8;
                xLf=xLs+xw+3*rD;//right of L
                xLW=xLf-xLs;//width of L
                xLl=xw;//length of left-line
                xLr=xLl;//length og right-line
                degree=231;
                g.drawArc(xLs,yL0-yh/2,xw,yh, 180, -degree);
                g.drawArc(xLs+rD,yL0-yh/2,xw,yh, degree, -degree-(degree-180));
                g.drawArc(xLs+2*rD,yL0-yh/2,xw,yh, degree, -degree-(degree-180));
                g.drawArc(xLs+3*rD,yL0-yh/2,xw,yh, degree, -degree);
                g.drawLine(xLs, yL0,xLs-xLl,yL0);
```

```
                g.drawLine(xLf, yL0,xLf+xLr,yL0);
//resistance
                int yR0,xRs,xRf,xRW,xRD,xRl,xRr;
                yR0=yL0;//y-center of R
                xRl=xLl;
                xRr=xLr;
                xRs=xLf+xLr+xRl+20;
                xRf=xRs+xLW;
                xRW=xRf-xRs;//width of R
                xRD=xRW/12;
                g.drawLine(xRs, yR0,xRs+xRD, yR0-yh/2);
                g.drawLine(xRs+xRD, yR0-yh/2,xRs+3*xRD, yR0+yh/2);
                g.drawLine(xRs+3*xRD, yR0+yh/2,xRs+5*xRD, yR0-yh/2);
                g.drawLine(xRs+5*xRD, yR0-yh/2,xRs+7*xRD, yR0+yh/2);
                g.drawLine(xRs+7*xRD, yR0+yh/2,xRs+9*xRD, yR0-yh/2);
                g.drawLine(xRs+9*xRD, yR0-yh/2,xRs+11*xRD, yR0+yh/2);
                g.drawLine(xRs+11*xRD, yR0+yh/2,xRs+12*xRD, yR0);
                g.drawLine(xRs, yR0,xRs-xRl,yR0);
                g.drawLine(xRs+12*xRD, yR0,xRf+xRr,yR0);
// capacitance
                int yC0,xCs,xCf,xCW,xCD,xCl,xCr;
                yC0=yL0;//y-center of C
                xCl=xLl;
                xCr=xLr;
                xCs=xRf+xRr+xCl+20;
                xCf=xCs+xRW;
                xCW=yh/5;//width of C
                g.drawLine(xCs-xCl, yC0,(xCs+xCf)/2-xCW/2, yC0);
                g.drawLine(xCf+xCr, yC0,(xCs+xCf)/2+xCW/2, yC0);
                g.drawLine((xCs+xCf)/2-xCW/2, yC0-yh/2,(xCs+xCf)/2-xCW/2, yC0+yh/2);
                g.drawLine((xCs+xCf)/2+xCW/2, yC0-yh/2,(xCs+xCf)/2+xCW/2, yC0+yh/2);

// print of values
                double r,c,el;
                r=50.0;
                c=50.0;
                el=10.0;
                float rf,cf,elf;
                rf=(float)r;
                cf=(float)c;
                elf=(float)el;
                String rs,cs,els;
                rs=Float.toString(rf);
                cs=Float.toString(cf);
                els=Float.toString(elf);
                g.drawString("R = "+rs+" (Ω)", (xRs+xRf)/2-35, yR0-yh/2-5);
                g.drawString("C = "+cs+" (μF)", (xCs+xCf)/2-35, yC0-yh/2-5);
                g.drawString("L = "+els+" (mH)", (xLs+xLf)/2-35, yL0-yh/2-5);

        }
}
```

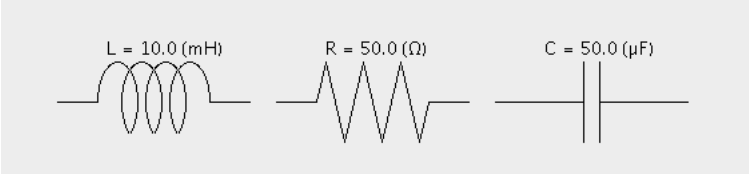


Figure 3: Basic print for the L, R, and C and their values, displayed using the Java program

In this basic numerical calculation involving a series circuit, six text fields for the resistance coefficient (R Ω), the capacitance coefficient (C μF), the inductance coefficient (L mH), the frequency of the applied voltage (f Hz), the time increments (h sec), and the number of calculations (n), are set on the display. For the calculations on parallel circuits, another text field for the series resistance (R_S Ω) is added. The inputted values for the R, C, and L are automatically displayed on each of the components, as shown in Figure 3. The calculated electric characteristics are plotted on the display, which is separated into two regions. The applied voltage $v(t)$, current $i(t)$, and charge $q(t)$ are plotted in the upper regions. Each of the voltages for the components, $v_R(t)$, $v_C(t)$, and $v_L(t)$ in the series circuit, or each of the currents of the components, $i_R(t)$, $i_C(t)$, and $i_L(t)$ in the parallel circuit are plotted in the lower regions.

4 RESULTS

The text fields are immediately presented on the display after the execution of the program by the appletviewer in Java. The calculation is initiated by clicking the start button after inputting values in the text fields. The figure of the circuit and the voltage-current characteristics are plotted immediately after the completion of the calculation. The time required to display the characteristics after clicking the start button is usually less than several seconds depending on the number of calculations and the performance of the computer. By changing the values in the text fields and re-clicking the start button, new characteristics can be obtained immediately.

4.1 Electric Characteristics of Series RCL Circuit

The typical AC characteristics of a series RCL circuit in which the resistance, capacitance, and inductance are comparably effective, that is, the impedance of each component is nearly equal, are shown in Figure 4. This figure shows the image on the display. The current is dominated by the resistivity, found by deducing the sum of the impedances of the capacitance and inductance, and the current results in same phase as the applied voltage. The phase of the charge is larger than that of the current by π/2 as shown in Equation (2). The phase of the voltage across the resistance is same as that of the current. On the other hand, the phase of the capacitance voltage is larger than the current by π/2 and that of the inductance voltage is less than the current by π/2. The phase of the capacitance voltage is π behind that of the inductance voltage. These basic electric characteristics of the R, C and L components are obtained from Equations (1) – (3), and can easily be shown visually on the display. By decreasing the value of C in the text field by a factor of 10, the new characteristics affected strongly by the capacitance, in which the phase of the current is less than that of the applied voltage by π/2 and the larger part of the applied voltage is applied to the capacitance, are obtained immediately, as shown in Figure 5. At a small t value, a transient characteristic appears, caused by the initial increase in the voltage from 0 to v_0. The increasing rate of the voltage is v_0/h and depends on the inverse of the increment of time. The rate is usually large and results in a large transient effect on the characteristics of the capacitance, as shown in

Figure 5. By increasing the value of R in the text field by a factor of 10, relative to that in Figure 4, the new characteristics affected strongly by the resistance can be obtained immediately, and by increasing that of L by a factor of 10, the new characteristics affected strongly by the inductance can also be obtained immediately. The maximum value of each variable and the final times are also shown on the display. By changing the values in the text fields and re-clicking the start button, the change of the electric characteristics in the series RCL circuit are easily and accurately recognized and are displayed immediately as an image on the display.

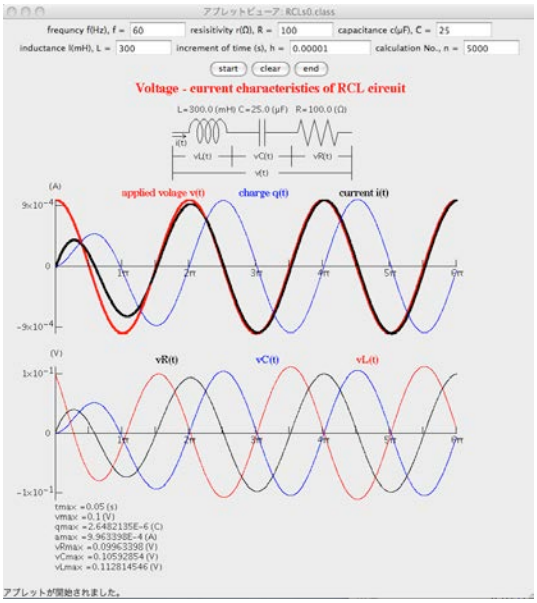


Figure 4: Typical AC characteristics of a series RCL circuit in which the R, C and L components are comparably effective. The balues used are shown in the text fields, f = 60 Hz, R = 100 Ω, C = 25 μF, and L = 300 mH. v(t), i(t), and q(t) are shown in the upper regions. $v_R(t)$, $v_C(t)$, and $v_L(t)$ are shown in the lower regions. The current results in similar phase to the applied voltage, and the phase of the voltage for each component is different by π/2 each other. The maximum values of the variables are also shown on the display

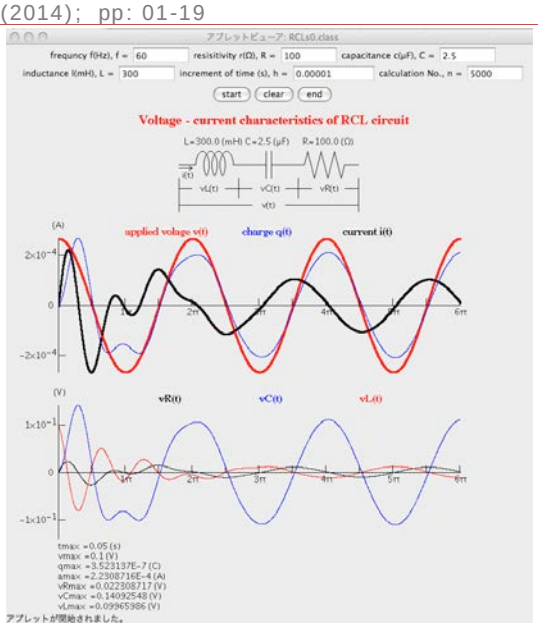


Figure 5: Updated AC characteristics of a series RCL circuit affected strongly by the capacitance after decreasing the value of C by a factor of 10 relative to that shown in Figure 4. The phase of the current is less than that of the applied voltage by π/2 and larger part of the applied voltage is applied to the capacitance

The typical DC characteristic, which corresponds to applying a rectangle voltage pulse, of the series RCL circuit in which the resistance, capacitance, and inductance are comparably effective are shown in Figure 6, using the same values in the text fields shown in Figure 4. By increasing the value of R in the text field by a factor of 10, relative to that shown in Figure 6, the new characteristics affected strongly by the resistance, in which the shape of the current is similar to that of the applied voltage and the larger part of the voltage is applied to the resistance except in the transient regions, can be obtained immediately, as shown in Figure 7. By increasing the value of L in the text field by a factor of 10 relative to that shown in Figure 6, the new characteristics affected strongly by the inductance, in which the shape of the current is shown as a integration of the applied voltage and the larger part of the voltage is applied to the inductance, can be obtained immediately, as shown in Figure 8. By decreasing the value of C by a factor of 10 and decreasing the value of L by a factor of 100, the new characteristics affected strongly by the capacitance, in which the shape of the current is shown as a differentiation of the applied voltage and the larger part of the voltage is applied to the capacitance, can be obtained immediately, as shown in Figure 9.

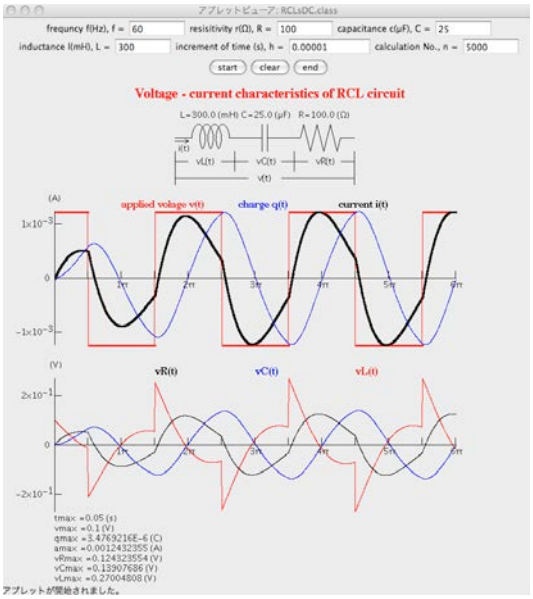


Figure 6: Typical DC characteristics of a series RCL circuit in which the R, C and L components are comparably effective. The values used are same as those shown in Figure 4

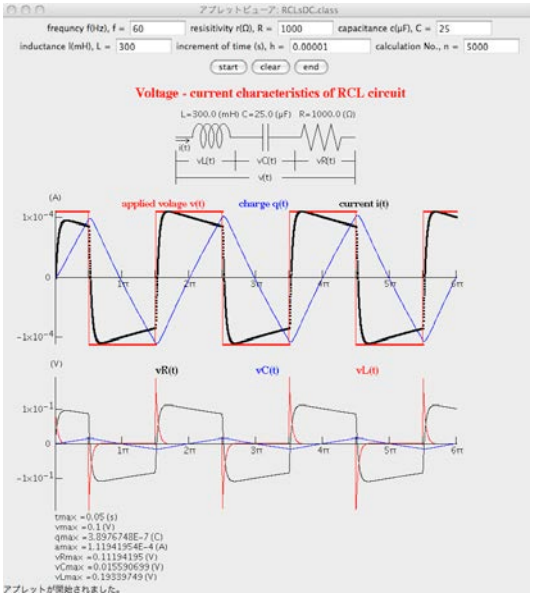


Figure 7: Updated DC characteristics of a series RCL circuit affected strongly by the resistance after increasing the value of R by a factor of 10 relative to that shown in Figure 6. The shape of the current is similar to that of the applied voltage

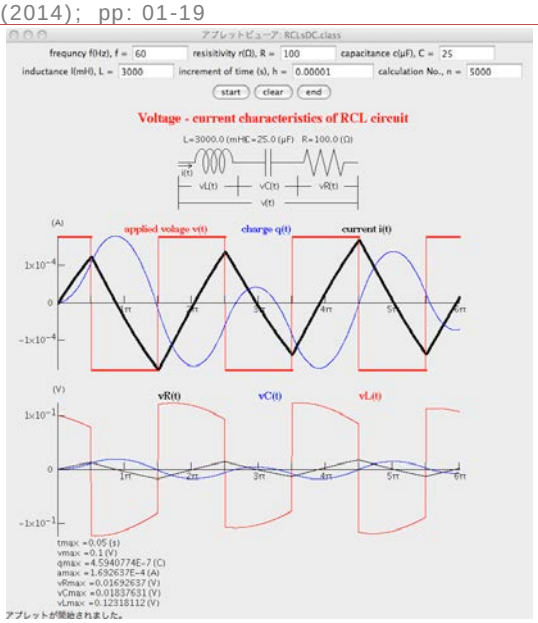


Figure 8: Updated DC characteristics of a series RCL circuit affected strongly by the inductance after increasing the value of L by a factor of 10 relative to that shown in Figure 6. The shape of the current is shown as an integration of the applied voltage

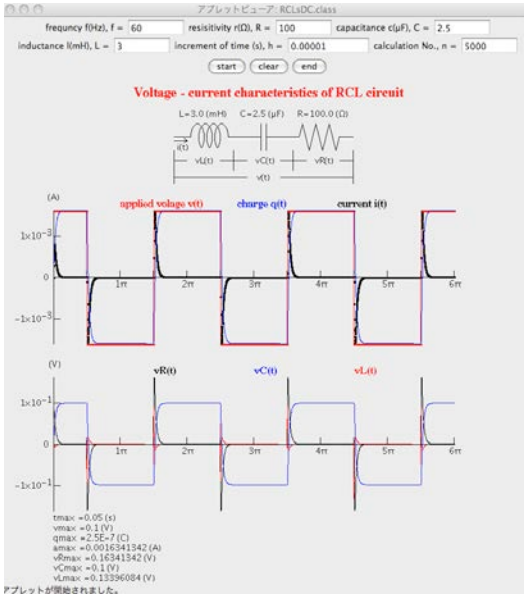


Figure 9: Updated DC characteristics of series RCL circuit affected strongly by the capacitance after decreasing the value of C by a factor of 10 and decreasing the value of L by a factor of 100 relative to the values shown in Figure 6. The shape of the current is shown as a differentiation of the applied voltage

4.2 Electric Characteristics of Parallel RCL Circuit

The typical AC characteristics of a parallel RCL circuit in which the resistance, capacitance, and inductance are comparably effective is shown in Figure 10. The used values in the text fields are same as the values used in Figure 4, except for the addition of R_S. The total current is dominated by the resistivity R, found by deducing the sum of the currents in the capacitance

and the inductance, and the total current results in the same phase as the applied voltage. The phase of the charge is same as that of the current, that is, it is in phase with the capacitance voltage. The phase of the resistance current is same as that of the voltage. On the other hand, the phase of the capacitance current is smaller than that of the resistance current by π/2, and that of the inductance current is larger than that of the resistance current by π/2. The phase of the capacitance current is π behind that of the inductance current. These basic electric characteristics of R, C, and L can be easily shown visually on the display. By increasing the value of C in the text field by a factor of 10, the new characteristics are affected strongly by the capacitance because the capacitive impedance decreases by a factor of 10 in comparison with the other impedances. Thus, the phase of total current is less than that of the applied voltage by nearly π/2, and the larger part of the total current is caused by the capacitance current. Theses characteristics can be obtained immediately, as shown in Figure 11. In this case, the initial transient current results in very large values, as shown in Figure 11. By decreasing the value of R in the text field by a factor of 10 relative to that in Figure 10, the new characteristics affected strongly by the resistance can be obtained immediately, and by decreasing that of L by a factor of 10, the new characteristics affected strongly by the inductance can also be obtained immediately. That is, by changing the values in the text fields and re-clicking the start button, the change in the characteristics of a parallel RCL circuit are recognized easily and accurately as an image showing the characteristics on the display.

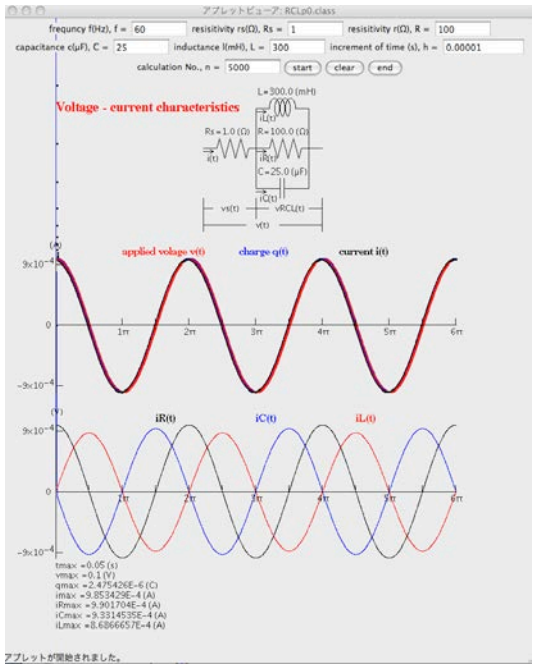


Figure 10: Typical AC characteristics of a parallel RCL circuit in which R, C and L components are comparably effective. The blues used are shown in the text fields, f = 60 Hz, R_S = 1 Ω, R = 100 Ω, C = 25 μF, and L = 300 mH. v(t), i(t), and q(t) are shown in the upper regions, and $i_R(t)$, $i_C(t)$, and $i_L(t)$ are shown in the lower regions. The total current results in similar phase to the applied voltage, and the phase of the current in each component is different by π/2 each other

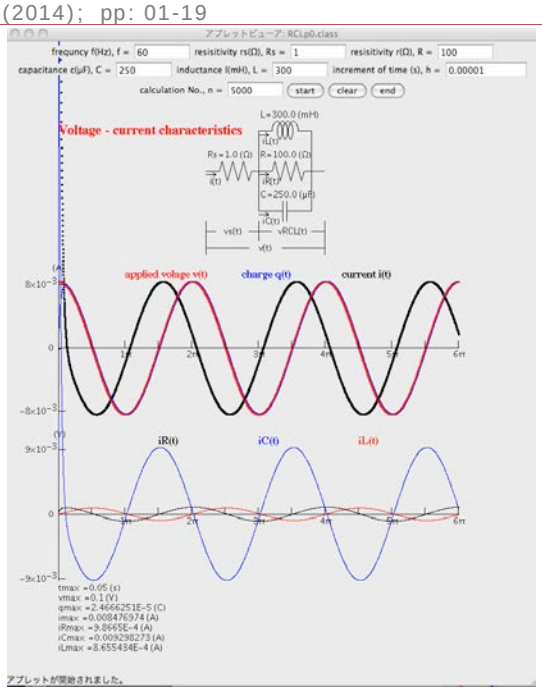


Figure 11: Updated AC characteristics of a parallel RCL circuit affected strongly by the capacitance after increasing the value of C by a factor of 10 relative to that shown in Figure 10. The phase of total current is less than that of the applied voltage by nearly π/2, and the larger part of the total current is caused by the capacitance current

The typical DC characteristics, which corresponds to applying a rectangle voltage pulse, of a parallel RCL circuit, in which the resistance, capacitance, and inductance are comparably effective, is shown in Figure 12 using same values in the text fields as the AC characteristics shown in Figure 10. In the parallel circuit, the transient current of the capacitance is very large at the time when the rectangle voltage is changing. The maximum transient current is inversely proportional to R_S in Figure 3 and is usually very large as shown in Figure 12 because of the small value of the R_S. The characteristics except in the transient regions are not obvious because the values are too small in contrast to the transient values. By decreasing the value of R in the text field by a factor of 10 relative to that shown in Figure 12, the new characteristics affected strongly by the resistance except in the transient regions, can be obtained immediately, as shown in Figure 13, in which the shape of the current is similar to that of the applied voltage and the larger part of the current is caused by the resistance current. By decreasing the value of L in the text field by a factor of 10 relative to that shown in Figure 12, the new characteristics affected strongly by the inductance except in the transient regions, can be obtained immediately, as shown in Figure 14, in which the shape of the current is shown as an integration of the voltage and the larger part of the current is caused by the inductance current. By increasing the value of C by a factor of 10 relative to that shown in Figure 12, the new characteristics affected strongly by the capacitance, in which the shape of the current is shown as a differentiation of the voltage and the larger part of the current is caused by the capacitance current, can be obtained immediately, as shown in Figure 15.

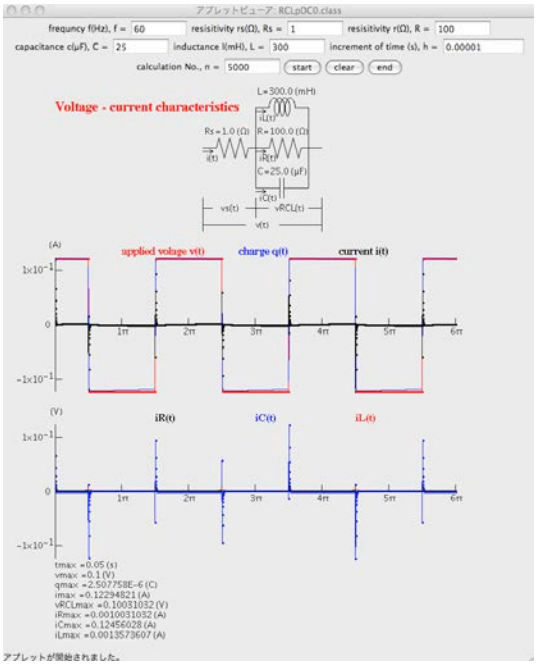


Figure 12: Typical DC characteristics of a parallel RCL circuit in which R, C and L components are comparably effective. The values used are same as those shown in Figure 10. The transient current of the capacitance is very large and the characteristics, except in the transient regions, are not obvious because the values are too small in contrast to the transient values

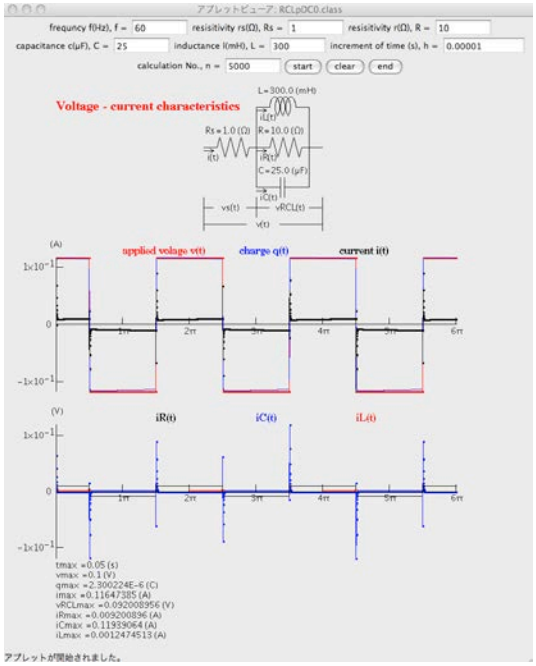


Figure 13: Updated DC characteristics of a parallel RCL circuit affected strongly by the resistance after decreasing the value of R by a factor of 10 relative to that shown in Figure 12. The shape of total current is similar to that of the voltage and the larger part of the current is caused by the resistance current, except in the transient regions

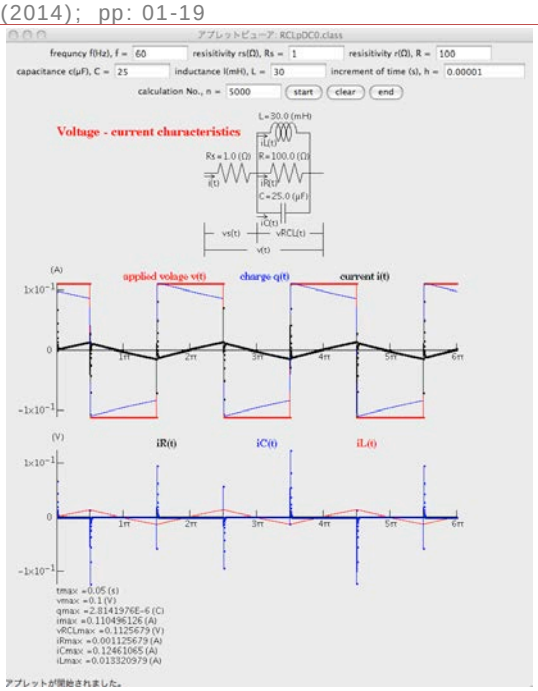


Figure 14: Updated DC characteristics of a parallel RCL circuit affected strongly by the inductance after decreasing the value of L by a factor of 10 relative to that shown in Figure 12. The shape of total current is shown as an integration of the voltage and the larger part of the current is caused by the inductance current, except in the transient regions

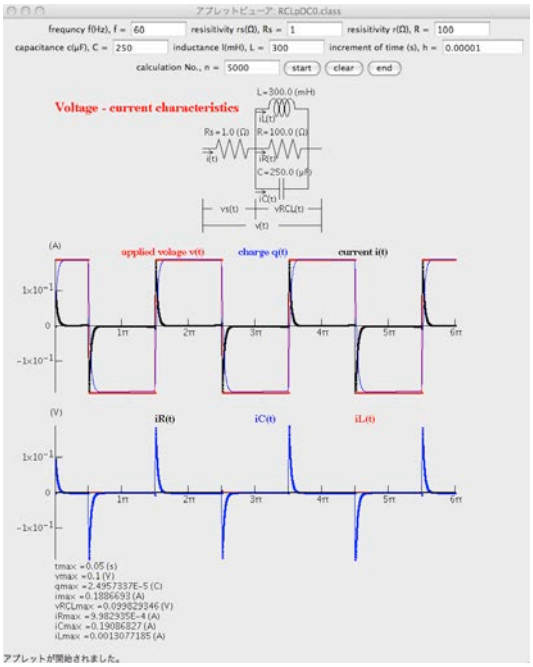


Figure 15: Updated DC characteristics of a parallel RCL circuit affected strongly by the capacitance after increasing the value of C by a factor of 10 relative to that shown in Figure 12. The shape of total current is shown as a differentiation of the voltage and the larger part of the current is caused by the capacitance current

5 DISCUSSION

The accuracy of the calculations using this program depend on the value of the time increment, h. If we use a too large h, the calculation is not carried out accurately and the electric characteristics cannot be obtained, as shown in Figure 16. The accuracy of the calculation is increased by using smaller h values. However, this leads to a larger number of calculations, thus a longer calculation time is needed to obtain the appropriate characteristics. It is better to choose a large h value by performing tentative calculations with a relatively small number of calculations.

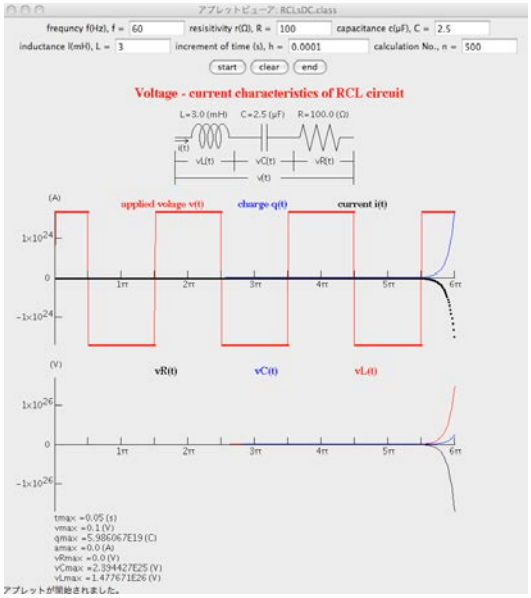


Figure 16: A sample of a non-accurate calculation using a large *h* value. The values used for the calculation are same as those in Figure 9, except for *h* = 0.0001 and *n* = 500

The transient current of the capacitance is very large at the time when the changes in the voltage are large, as mentioned above. This results in the currents in the other components becoming relatively small such that their characteristics cannot be clearly presented, as shown in Figure 12. In this case, the characteristics can be clearly presented by selecting the maximum and minimum points on the display from the currents except in the transient current, as shown in Figure 17.

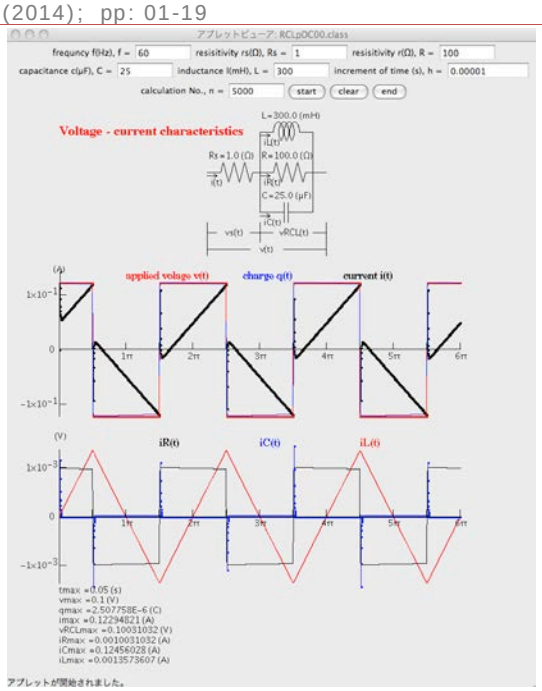


Figure 17: Selection of the maximum and minimum points on the display from the currents except in the transient current in the capacitance. The values used for the calculation are same as those in Figure 12

A typical oscillation of current in a series LC circuit applied a DC voltage is shown in Figure 18. The observed frequency of the oscillation, 318 Hz, agrees with the theoretical value obtained using $1/(2\pi\sqrt{LC})$.

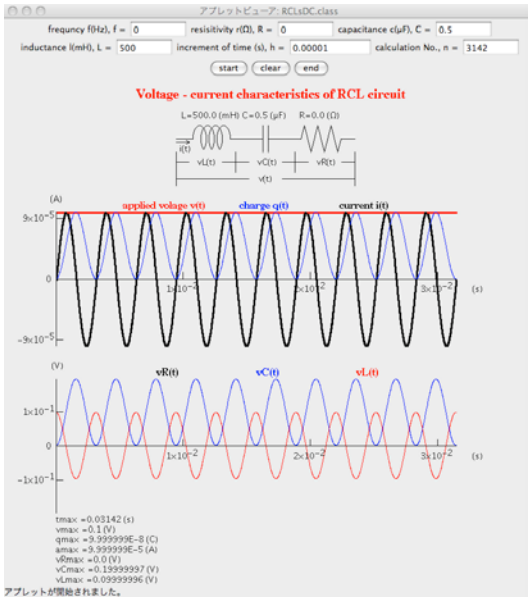


Figure 18: The typical oscillations of the current in a series LC circuit with a DC voltage applied

6 CONCLUSION

Java programs in a GUI environment have been developed to simulate the electric characteristics of a resistance, a capacitance, a inductance, and their circuits. The values of the

selected parameters for the numerical calculation are set using the text fields on the display, and the calculation is initiated by clicking the start button after inputting these values. The calculated results are plotted immediately after the completion of the calculation as the figures of the voltage and current on the display. By changing the values in the text fields and re-clicking the start button, the new results can be displayed immediately. The simulations of the characteristics depending on the values of each component in the electric circuit can be obtained easily and accurately as the changes of the voltage and current with time on the display. The time required to run a simulation is very short, less than several seconds using a personal computer. These Java programs are useful in education applications for image learning about the characteristics of electric circuits because of their ability to quickly provide accurate depictions of the fundamentals of electric circuits.

This Java program can be applied for rapidly and accurately image learning about the phenomena expressed by ordinary differential equations, because the Runge-Kutta method is used for the calculations. By using the Crank-Nicolson's implicit method and the Gauss-Seidel's iteration method, these Java simulations are also useful in education applications for rapidly and accurately image learning about complicated phenomena expressed by partial differential equations such as the diffusions of atoms [2] and the propagations of heat and wave.

REFERENCES

[1]. M. Morooka, S. Qian, and M. Morooka, *Image Learnig of Charge Motion in Electric and Magnetic Fields by Java Programming*, Transactions on Machine Learnig and Artificial Intelligence, 2014. **2**(2): p.1- 19.

[2]. M. Morooka, *Java Simulation of Au Diffusion in Si Affected by Vacancies and Self-Interstitials: Partial Differential Equations*. Journal of Sofware Engineering and Applications, 2012. **5**(10): p. 764-776.

TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 2, ISSUE 3
ISSN: 2054 - 7390

Unsupervised Machine Learning Techniques for Detecting Malware Applications in Wireless Devices

[*]Jackson Akpojaro[1], Princewill Aigbe[1], Ugochukwu Onwudebelu[2]

[1]Department of Mathematics and Computer Science, Western Delta University, Oghara, Delta Stat, Nigeria
[2]Department of Mathematics & Computer Science, Federal University, Ndufu, Alike Ikwo, Abakiliki, Ebonyi State,Nigeria

* jakpojaro@yahoo.com

ABSTRACT

It is no doubt that we are in the era of 'big data', and different machines and tools are being developed every day to enable users to effectively access, manipulate and process data to provide timely information needed for decision making. The situation has led to increasingly use of wireless devices including smartphones, tablets, pacemakers, etc., with different platforms. As professionals including doctors, engineers, scientists, artists, etc., use these devices in accessing, process and disseminating information services are available, so also malware attackers are strategizing. Hence the last one decade has witnessed constant literatures in the design and development of both supervised and unsupervised machine learning algorithms to checkmate malware applications in wireless devices. In this paper, we study the properties of unsupervised learning algorithms; in particular, we quantify the performance of these algorithms under two scenarios; using data sets from unknown attackers and data sets from known attackers. Our findings show that the recently γ-algorithm appears superior to the other unsupervised algorithms investigated.

Keywords: big data, wireless devices, malware, supervised algorithms, unsupervised algorithms.

1 INTRODUCTION

The use of wireless devices such as smartphones, tablets, pacemakers, etc. have become very popular among professionals because they provide convenience and easy access to timely information. As the functionalities and capabilities of these devices are increasing rapidly within

DOI: 10.14738/tmlai.23.206
Publication Date: 9th June 2014
URL: http://dx.doi.org/10.14738/tmlai.23.206

a short space of time with every new model, health experts and other users are beginning to rely on them to conduct diagnoses, businesses, interact with families and friends, play games, shopping, etc. Medical scientists have keyed in into this technology, using smartphones and wireless pacemakers for diagnoses, early testing, and electronic medication alerts with the aim of reducing prescribing errors [20]. A pacemaker is a small device that is placed in the chest or abdomen to help control abnormal heart rhythms, while the recently developed mobile phone application could help make monitoring conditions such as diabetes, kidney disease, and urinary tract infections much clearer and easier for both patients and health professionals, and could be used to slow or limit the spread of pandemics in the developing world [20].

As the technology is developing rapidly with increasing applications, so also security threats that target these applications are on the increase. In fact, malicious users and hackers are taking advantage of lack of standard security mechanisms to design mobile-specific malware that can access sensitive data, steal users' phone credit, or deny users' access to key functionalities in the device [18]. In the Juniper networks report on mobile threats, malware attacks have increased by 155 % across all platforms. In particular, devices with android platform had the highest malware growth rate [19].

To mitigate these security threats, the last one decade has witnessed a constant stream of literature on design and development of machine learning algorithms to detect malware in wireless devices. In this paper, we evaluate the performance of some the proposed and currently used unsupervised algorithms. In particular, we study their properties and characterize their performance under two scenarios: data sets from known attack and data sets from unknown attacks.

Summarizing, our main findings in this paper are:

- We study the properties of some unsupervised learning algorithms.
- We create different data sets and run the algorithms to produce experimental results.
- We find that the recently proposed γ-algorithm demonstrates some significant performance difference in both data sets with known attacks and data sets with unknown attacks.
- γ-algorithm is shown to be more promising than other unsupervised algorithms evaluated.

The reminder of this paper is organized as follows: Section 2 reviews some relevant background work. Section 3 discusses types of machine learning, while Section 4 describes the unsupervised algorithms we have evaluated. Section 5 presents our experimental results, while the results are discussed in Section 6. The paper is concluded in Section 7 with proposed research direction to formalize probabilistic models to quantify currently used supervised and unsupervised algorithms in static and dynamic environments with a view to determining allocation of scarce resources to promising algorithms at early design stage.

2 RELATED WORK

It is no doubt that we are in the era of 'big data', and different machines and tools are being developed to ensure that users have access to timely information to make decisions wherever they are. Many professionals have keyed in, doctors and other health experts use smartphones and other wireless devices to conduct medical diagnoses and tests. As these wireless devices with different operating platforms are increasing, developers of malware are strategizing. This has intensified and motivated research in machine learning to checkmate malware in different platforms. More heuristic methods have been proposed in this field to tackle specific problems. For instance, neural network models [21] have been inspired by the support vector classifiers [22, 23, 24]. Weston et al. [25] focused on the study of outliers from the perspective of the classification problem.

In the last decade, the field of semi-definite programming (SDP) has opened windows of opportunities for designing promising machine learning techniques. The consistency of researchers in this field has yielded a viable technology with efficient characteristics similar to quadratic programming [26]. Lanckreit et al. [27] demonstrated how SDP is used to optimize the kernel matrix for a supervised support vector machine (SVM). Xu et al. [28], De Bie et al. [29] developed new unsupervised and semi-supervised training techniques for SVMs based on SDP.

Several machine learning techniques have been applied for classifying applications with focus on detecting malware [30, 31]. Their goal is to classify applications into two main categories; malware or goodware. In [32, 33], the authors tried to classify applications by specifying the malware class (e.g., worms, Trojan, virus, etc.).

As the number of malware samples is exponentially increasing, particularly with Android platform, several techniques have been proposed to tackle the surge. Shabtai et al. [34] trained machine learning models, e.g., parsing *apk* which contains *xml* and counting xml elements, attributes or namespaces. They evaluated their model using information gained, fisher score, and Chi-square. They obtained 89% of accuracy classifying applications into two categories: tools and games. Recently, the γ-algorithm was proposed [11]. It is a graph-based outlier which assigns to every example the γ-score, which is the mean distance to the example k-nearest neighbors. Our experimental results show that this algorithm appears superior to other unsupervised algorithms in detecting malware in data sets involving known and unknown attackers.

The surveyed works provide the background for this paper. We study the properties of some unsupervised learning algorithms. In particular, we evaluate their performance under two scenarios: we create data sets with known attackers and data sets with unknown attackers. We

run the algorithms under these two situations and find that the γ-algorithm is more promising in detecting malware than the other algorithms investigated.

3 MACHINE LEARNING

Machine learning is a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict futures data, or to perform other kinds of decision making under uncertainty, for example, planning how to collect more data. Machine learning has been an active research area for more than a decade with focus on design and development of new algorithms that allow the computers to think and decide based on data [1].

Machine learning usually distinguishes three cases: supervised, unsupervised, and reinforcement learning. In supervised learning approach, the goal is to learn a mapping from input χ to output y, given a labeled set of input-output pairs [2], which is defined by a learning function,

$$\partial = \{(x_i, y_i)\}_{i=1}^{n} \quad (1)$$

where ∂ is the training data set, and n is the number of training examples. In its simplest form, each training inputs x_i could be features, attributes or covariates. More generally however, x_i could be a complex structured object, e.g., an image, email message, segment of application, sentence, etc. Variants of supervised learning algorithms include Bayesian Networks [3], Decision Trees [4], k-Nearest Neighbor (KNN) [5], and Support Vector Machine (SVM) [6].

In unsupervised learning problems, we have unlabeled inputs with a learning function defined as,

$$\partial = \{x_i\}_{i=1}^{n} \quad (2)$$

The aim is to find (or discover) interesting patterns or structures in the data set that can help to make informed decisions. In a purely unsupervised learning problem, agent cannot learn what to do because it has no specific output information as to what constitutes a correct action or a desirable state [7].

In reinforcement learning, rather than being told what to do, a reinforcement agent learns how to act or behave when given occasional reward or punishment signals. It is the most general of the three categories. The following subsection reviews unsupervised machine learning algorithms, explore how unlabeled data are clustered with a view to revealing some hidden structures to detect malware applications in wireless devices.

4 UNSUPERVISED ALGORITHMS

In unsupervised malware detection problems, we receive a large data set (e.g., emails) which contains both normal and buried malicious data within the data set [8]. Unsupervised algorithms have general features of able to process unlabeled data to detect malicious data that otherwise could not have been detected. In particular, some of these algorithms can automate the manual audit of data in forensic analysis by assisting analysts to focus on the suspicious elements in the data.

Unsupervised malware detection algorithms make two specific assumptions about the received data set: first, the number of normal instances outnumbers the number of malware instances. Secondly, the malicious instances are qualitatively different from the normal instances. Since the malware instances are both different from the normal instances and rare, they will appear as outliers in data set, which can be detected. In the light of this, we discuss the following unsupervised algorithms we have implemented in this work.

k-Means Clustering: The k-mean clustering algorithm is a variant of the partition clustering technique. It is a classical algorithm [9]. Its methodology is that after an initial random assignment to example k clusters, the centers of clusters are computed and examples are assigned to the clusters with the closest centers. This followed with several iterations until the cluster centers do not significantly change. Once the cluster assignment is fixed, the mean distance of an example to clusters is used as the score. There are simple approximations that speed up this algorithm considerably. For instance, one can project the data set and make cuts along selected axes, instead of using the arbitrary hyperplane divisions that are implied by choosing the nearest cluster center [10]. Details of how to speed up things are found in [10].

γ-Algorithm: The γ-algorithm [11] proposed recently is a graph-based outlier which assigns to every example the γ-score, which is the mean distance to the example k-nearest neighbors. It ignores the distances to the closer neighbors. More formally, a refined index that takes the distances to all k nearest neighbors is given thus [11];

$$\gamma(x) = \frac{1}{k}\sum_{j=1}^{k} \| x - z_j(x) \| \qquad (3)$$

where $\gamma(x)$ is x's average distance to its k nearest neighbors, $z_1(x),\ldots, z_j(x) \in \{x_{1,\ldots}, x_j\} \subset \Re^d$ (where $\Re^d$ refers to d-dimensional Euclidean space).

Divisive Hierarchical Clustering (DHC) - top-down: The divisive hierarchical clustering [12, 13] starts with one cluster of data set and each iteration split the most appropriate cluster until a stopping criterion such as a requested number k of clusters is achieved. Its implementation is described in [14].

Agglomerative Hierarchical Clustering (AHC) - bottom-up: An alternative to the top-down method for forming a hierarchical structure of clusters is the bottom-up approach called agglomerative clustering. This idea was proposed many years ago and has recently enjoyed a resurgence in popularity [10]. It starts with each data set in a separate cluster and at each iteration it merges the most similar clusters until the stopping criterion is met. Agglomerative clustering algorithms are categorized as single-linkage, complete-linkage, and average-linkage algorithms depending on the method each defines inter-cluster similarity.

The single-linkage algorithm defines the minimum distance between two clusters – the distance between their two closest numbers [10]. That is, it defines the similarity of two clusters C_i and C_j as the similarity of the least similar data $D_i \in C_i$ and $D_j \in C_j$ as;

$$S_{sk}(C_i, C_j) = \min_{D_i \in C_i, D_j \in C_j} |\cos(D_i, D_j| \quad (4)$$

where S refers to similarity and sk is single-linkage. Since this measure takes into account only the two closest members of a pair of clusters, the procedure is sensitive to outliers; the addition of a single new instance can radically alter the entire clustering structure.

The complete-linkage algorithm measures the maximum distance between the clusters. Two clusters are considered close only if all instances in their union are relatively similar. More formally, it defines the similarity of two clusters C_i and C_j as the similarity of the two most similar data $D_i \in Ci$ and $D_j \in C_j$ as;

$$S_{ck}(C_i, C_j) = \max_{D_i \in C_i, D_j \in C_j} |\cos(D_i, D_j| \quad (5)$$

where ck refers to complete-linkage. This measure which is also sensitive to outliers seeks compact clusters with small diameters. However, some instances may end up much closer to other clusters than they are to the rest of their own cluster.

The average-linkage algorithm is a measure which tries to avoid the problem inherent in centroid-linkage method since centroids are not instances and the similarity between them may be impossible to define. The average-linkage method defines the similarity of two data C_i and C_j as the average of pairwise similarities of the data from each cluster as;

$$S_{ak}(C_i, C_j) = \frac{\sum_{D_i \in C_i, D_j \in C_j} |\cos(D_i, D_j)|}{n_i n_j} \quad (6)$$

where ak is average-linkage, n_i and n_j are sizes of clusters C_i and C_j respectively.

Quarter-sphere Support Vector Machine (QSSVM): The quarter-sphere SVM [15] detects malicious data based on the idea of fitting a sphere onto the center of mass of data. An anomaly score is defined by the distance of a data point from the center of the sphere.

Choosing a threshold for the attack scores determines the radius of the sphere enclosing normal data points.

5 EXPERIMENTAL RESULTS

In this section, we evaluate the tradeoffs of the unsupervised learning algorithms briefly reviewed in Section 4. We evaluate the algorithms under two scenarios; first, we assume that the training and test data come from unknown attacks. Under the second scenario, we violated this assumption by taking data sets in which attacks unseen in training data are present in test data. Based on these, we created 6 data sets 200, 300, 400, 500, 1000, 2000 android applications (see Table 1). First we extract the necessary features from the applications to identify known malware (e.g., Adware, worm, Trojan, virus, rootkit, etc.), while in the second case, we pretend that the data sets contained malicious and normal data without classification.

Table 1: Datasets

Data Set #	No. of Samples	No of Features
1	200	120
2	300	145
3	400	148
4	500	175
5	1000	250
6	2000	318

We find that as the number of the samples increases, the performance difference of the algorithms becomes slightly significant. Hence we chose to provide the experimental results of the data set with 2000 samples (see Table 3 and Table 4). The evaluation metrics, true positive ratio (TPR), false positive ration (FPR), accuracy, and area under the ROC curve (AUC) are formalized and discussed in [16]. We use these formulae to obtain our experimental results as shown in Table 2 and Table 3 respectively.

Table 2: Obtained result for known attacks.

Algorithm	TPR	FPR	AUC	Accuracy (%)
γ -algorithm	0.96	0.10	0.98	98.33%
k -Means Clustering	0.90	0.08	0.86	91.12%
DHC	0.91	0.11	0.92	91.01%
AHC (single-linkage)	0.93	0.09	0.93	93.04%
QSSVM	0.89	0.19	0.85	91.11%

Table 3: Obtained result for unknown attacks

Algorithm	TPR	FPR	AUC	Accuracy (%)
γ -algorithm	0.98	0.08	0.99	99.54%
k -Means Clustering	0.89	0.10	0.86	91.11%
DHC	0.91	0.11	0.92	91.12%
AHC (single-linkage)	0.93	0.11	0.93	93.04%
QSSVM	0.90	0.19	0.85	91.11%

6 DISCUSSIONS

As presented in Table 2 and 3, the algorithms exhibit no significant difference in performance between known and unknown attacks except the γ-algorithm. This is because the two data sets differ merely in the set of attacks contained in them. However, only the γ-algorithm is shown to be promising in both data sets. It is 98% (FPR) better in detecting malware for unknown attacks as against 96% (FPR) for data sets containing known attacks. More generally, the γ-algorithm is not only significant in performance (accuracy (%)) in both data sets, but it also better than the other algorithms tested. The k-means clustering has the least TPR of 0.89 (see Table 3), but compares favorably with the QSSVM algorithm. Our results corroborate the work of Borja Sanz et al., [16], Pavel Laskov et al. [17], and Stafan Harmeling et al. [11].

The limitation of our results is that they are based on 2000 samples. We believe more significant performance differences among the algorithms could be revealed in larger samples (e.g., between 100,000 and more) that require more computational time and other resources. For brevity, we leave this investigation to others.

7 CONCLUSIONS AND FUTURE WORK

We have presented an experimental framework in which the unsupervised learning algorithms are evaluated in detection of malware in wireless devices. Our experimental results demonstrate no major significant performance difference in both unknown and known data sets except the γ-algorithm, which is not only superior to the other algorithms but also exhibits performance difference in both data sets. We find that as the data sets get larger, all the algorithms exhibit some performance differences; hence we chose to present the results for 2000 samples. We believe that with larger samples, e.g., 100,000 or more data sets, the algorithms would exhibit more significant results that could be further used to characterize them.

In our future research, we plan to formalize analytic model to quantify both supervised and unsupervised learning methods using common metric(s). In particular, we will analyze and evaluate these algorithms in both static and dynamic environments. In the dynamic scenario, we plan to introduce probabilistic models to enable us determine in real time the relative performance of these algorithms in detecting malware and also measure what new (or upgraded) algorithms claimed to be contributing at development stage. In doing so, scarce resources could be channeled to only new algorithms that demonstrate promising contribution to the current state of the art in machine learning.

REFERENCES

[1]. Bishop, C., Pattern recognition and machine learning, Springer New York, 2006.

[2]. Murphy, K., Machine learning: A Probabilistic perspective, MIT Press, Cambridge, MA, 2012.

[3]. Pearl, J., Reverend Bayes on inference engines: A distributed hierarchical approach, In Proceedings of National Conference on Artificial Intelligence, 1982, pp. 133-136.

[4]. Quinlan, J., Induction of decision trees, Machine learning 1(1), 1986, pp. 81-106.

[5]. Fix, E., Hodges, J. L., Discriminatory analysis: Nonparametric discrimination: Small sample Performance, Technical Report Project 21-49-004, Report number 11, 1952.

[6]. Vapnik, V., The nature of statistical learning theory, Springer, 2000.

[7]. Russell, S. and Norving, P., Artificial Intelligence: A Modern approach, 2nd Edition, Prentice Hall, Upper Saddle River, New Jersey 07458, 2003.

[8]. Leonid, P., Leazar, E., and Salvatore, J., Instruction Detection with unlabeled Data using Clustering. In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA 2001) Philadelphia, PA, 2001.

[9]. Duda, R., Hart, P., and Stork, D., Pattern Classification, Second Edition, John Willey & Sons, 2001.

[10]. Ian, H., Eibe, F., and Hall, M., Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, Morgan Kaufman Publishers, Burlington, MA 01803, USA, 2011.

[11]. Harmeling, S., Dornhege, G., Tax, D., Meinecke, F., and Miller, K., From Outliers to Prototypes: Ordering Data, Neurocomputing Vol. 69, pp. 1608-1618, 2006.

[12]. Jain, A., Murty, M., and Flynn, P., Data clustering: A review, ACM Computing Surveys, Vol. 31, No. 3, pp. 264–323, September 1999.

[13]. Berkhin, P., Survey of clustering data mining techniques, Research paper, Accrue Software, http://www.accrue.com/products/researchpapers.html, 2002.

[14]. Kaufman, L. and Rousseeuw, P., Finding groups in data, Wiley, New York, NY, 1990.

[15]. Laskov, P., Schafer, C., and Kotenko, I., Intrusion Detection in Unlabeled Data with Quarter-sphere Support Vector Machines. In proceedings DIMVA, pp. 71-82, 2004.

[16]. Borja, S., Igor, S., Javier, N., Carlos, L., Inigo, A., and Pablo, G., MADS: Malicious Android Applications Detection through String Analysis. Lecture Notes in Computer Science, Vol. 7873, pp. 178-191, 2013.

[17]. Laskov, P., Diissel, P., Schafer, C., and Rieck, K., Learning Instruction Detection: Supervised or Unsupervised?. Fraunhofer-FIRST IDA, 12489 Berlin, Germany, 2006.

[18]. Zami, A., and Zawi, W., Permission-Based Android Malware Detection. International Journal of Scientific and Technology Research, Vol. 2, Issue 3, pp. 228-234, 2013.

[19]. Juniper Networks: 2011 Mobile threats report, February 2012.

[20]. Muanya, C., Smartphones, wireless pacemakers, turned into portable medical devices. The Guardian, p.31, Thursday, March 27, 2014.

[21]. Marshland, S., Online novelty detection through self-organization with application to inspection robots. Ph.D. Thesis, University of Manchester, 2001.

[22]. Scho, B., J. Shawe-Taylor, P., Smola, A., and Williamson, R., Estimating the support of a high-dimensional distribution, Neural Computation Vol. 13 Issue 7, pp. 1443-1471, 2001.

[23]. Campbell, C., and Bennett, K., A linear programming approach to novelty detection. Advances in Neural Information Processing Systems, Vol. 13, MIT Press, Cambridge, MA, pp. 395-401, 2001.

[24]. Tax, D., and Duin, R., Uniform object generation for optimizing one-class classifiers, J. Mach. Learn. Research, pp. 155-173, 2001.

[25]. Weston, J., Chapelle, O., and Guyon, I., Data cleaning algorithms with applications to micro-array experiments, Technical Report, BIOwulf Technologies, 2001.

[26]. Boyd, S., and Vandenberghe, L., Convex Optimization, Cambridge, U. Press, 2004.

[27]. Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L., and Jordan, M., Learning the kernel matrix with semidefinite programming, Journal of Machine Learning Research, 2004.

[28]. Xu, L.; Neufeld, J.; Larson, B.; and Schuurmans, D., Maximum margin clustering. In Advances in Neural Information Processing Systems 17 (NIPS-04), 2004.

[29]. De Bie, T., and Cristianini, N., Convex methods for transduction, In Advances in Neural Information Processing, 16 (NIPS-03), 2003.

[30]. Santos, I., Laorden, C., and Bringas, P., Collective classification for unknown malware detection, In Proceedings of the 6th International Conference on Security and cryptography (SECRYPT), 2011.

[31]. Y. Ye, Y., Wang, D., Li, T., and Ye, D., IMDS: Intelligent malware detection system, In Proceedings of the 13th ACM SIGKDD International conference on Knowledge discovery and data mining, ACM, pp. 1043-1047, 2007.

[32]. Rieck, K., Holz, T., Willems, C., Dussel, P., and Laskov, P., Learning and classification of malware behavior, In Proceedings of the 2008 Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA), pp. 108-125, 2008.

[33]. Tian, R., Batten, L., Islam, R., and Versteeg, S., An automated classification system based on the strings of trojan and virus families, In Malicious and Unwanted Software MALWARE), 2009 4th International Conference on IEEE, pp. 23-30, 2009.

[34]. Shabtai, A., Fledel, Y., and Elovici, Y., Automated Static Code analysis for classifying Android applications using machine learning," 2010 International Conference on Computational Intelligence and Security, pp. 329–333, 2010.

TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 2, ISSUE 3
ISSN: 2054 - 7390

The Artificial Intelligence Development Axioms (A.I.D.A.)

N. Aljaddou
Department of Physics, University of Nebraska at Omaha, Omaha (NE) 68182
naljaddou@unomaha.edu

ABSTRACT

Within this paper is a set of critical new developmental principles concerning the inevitable historical evolution of the faculties of artificial intelligence, as well as the humans manipulating their capabilities. The results may be referred to as *axioms* as they are contingent on irreducible mathematical models which map the capacities of said artificial intelligence, and the game theoretic considerations of the optimal decision-making of their sentient counterparts. The paper is divided into four primary sections, covering the four primary principles of A.I.D.A., with additional preliminary and concluding sections. It is to be stressed that these are *inevitable* principles of artificial intelligence development, not merely hypothetical considerations, and this fact emphasizes the importance of their acknowledgment and dissemination within the scope of current academia and scholastic discourse. This paper, and the development of artificial intelligence research, is indebted to the work of many great minds in the century past; however the most prominent figure in whose name this work is dedicated, is the great Hungarian-American mathematician and technologist, John Von Neumann, who first coined the term "technological singularity", of which this work is the precise elaboration.

Keywords: Weak A.I., Critical Technological Capacity Point, Critical Governance Point, Von Neumann Sphere, Orbisphere

1 INTRODUCTION

"Some people say that computers can never show true intelligence whatever that may be. But it seems to me that if very complicated chemical molecules can operate in humans to make them intelligent then equally complicated electronic circuits can also make computers act in an intelligent way. And if they are intelligent they can presumably design computers that have even greater complexity and intelligence."

DOI: 10.14738/tmlai.23.190
Publication Date: 9th June 2014
URL: http://dx.doi.org/10.14738/tmlai.23.190

- Dreyfus, Hubert L.; Dreyfus, Stuart E. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer* New York: Free Press (2000)

2 PRELIMINARY CONCEPTS

Strong A.I. vs. Weak A.I.

The benefits of developing weak A.I. (program contingent) automated systems far outrank strong A.I. (independent mechanisms which could potentially host an independent operating brain – IOB – which would most likely have to be achieved through the design of synthetic neuronal synapses). Strong A.I. raises power and ethics concerns; weak A.I. does not, and is functionally more efficient as it may be run on a quantum computer system. In the end, it may have the capacity to fully model consciousness from a brain scan, but it can only simulate consciousness – never achieving it.

The A.I. Signature Capacity

That which defines A.I. as being distinct from a sheer calculating machine is its capacity to calculate its own calculations (initial assessments from observation, according to its operating system). This gives it broad binary decision-making capabilities which enable it to edit its future functionality. It may be theoretically described as a computational bijection matrix, which maps one series of a data set to a recurrence-functional output, perpetually modifying its subsequent operations. With this being the case, it can "improve" its very own functionality eventually – beyond even the capacity of its initial mechanical parameters. Herein is born the first principle.

3 THE CRITICAL TECHNOLOGICAL CAPACITY POINT (CTCP)

The point at which human input in technological progress has been alleviated by a sufficiently advanced artificial intelligence which can design increasingly advanced artificial intelligences in a recursive manner, having the capabilities to address and design all auxiliary technological needs and concerns (in an optimal fashion). The result of CTCP is called ATE (Automated Technological Evolution).

Modus Operandi

The self-editing capacities of the employed weak A.I. will not only allow them to self-improve, but to design in practice a means for their systems to be improved beyond the current physical mechanisms in place. This will at first take the form of displayed schematics, however, as this process is recursively generated, the prototype following the first few iterations will likely be endowed with much more efficient construction capabilities, and will be able to carry out precision crafting of the A.I. system which will eventually replace it. In theory, the A.I. being produced will become so advanced that maximally efficient quantum computational operating systems and optimization of this procedure will be reached relatively shortly – as this process is

exponentially exponential (due to the bijective editorial process intrinsic in the programming) this may in principle be achieved in as few as ten iterations.

Another way of looking at it is if humans themselves are capable of designing a quantum computer in the near future now as it stands, a sufficiently advanced A.I. could do so after very few self-recreating recursions. Their mechanical construction precision (even of their construction of constructive equipment itself) will far outrank any human or normal computer capability, and it should reach the level of molecular and subatomic manipulation after the first little iteration, provided the prototype was sufficiently advanced. After this process is complete, total ATE will have been reached, without the need for further human intervention, resulting in Technological Optimization Capacity (TOC), at which point the processing power of the A.I. will have reached Bremermann's Limit of computational ability.

Form and Function

It is likely that the notion of a stable structure machine will be a thing of the past with the newly introduced micro-subset nanotechnologies available, which will render any given device continuously adaptably self-mutating. Additionally, intense magnetic fields will have the ability to manipulate series of these "microprobes" into fluidic states which can reassemble into arbitrary solid structures. Needless to say, the potentialities for manipulation of these technologies is limitless, and could unfortunately possibly be used to cause the greatest calculable damage to a human populace. Leading to the next principle.

4 THE CRITICAL GOVERNANCE POINT (CGP)

The point at which human government becomes arbitrarily classified, data-collecting, and controlling, in conjunction with the achievement of the aforementioned CTCP, for necessity of guarding the unlimited manufacturing capability of the acquired artificial intelligences (which could be used for weapons-producing purposes).

The Aim of Government

The task of government, fundamentally, is to reduce as much risk to the species as possible. There is literally no greater risk posed to the public at large than the development of advanced A.I. capabilities. Analytic think tank members will have envisioned the cost-benefit implications of the development of the modern computer and its offshoots so thoroughly, that they will implement programs to develop A.I. in advance of the general public (a la the Manhattan Project for the atom bomb). This is merely an applied solution of the Nash Equilibrium in the appropriate scenario.

Government's Solution to the A.I. Dilemma

The government will see the only available precautionary prescription as achievement of CTCP themselves and mitigation of its development by the public through advanced monitoring using the newly developed A.I.

5 THE VON NEUMANN SPHERE

The ultimate fruit of the combination of the two critical points is the *Von Neumann Sphere* (analogous to the Dyson Sphere, although surrounding only the earth, and named after the inventor of the modern computer and coiner of the term "technological singularity"), a multitudinous, interlinked, geosynchronously orbiting network of artificial intelligence satellites monitoring all human activity on varying electromagnetic frequencies, collecting all available data, from ostensible superficialities to the very thought processes of citizens from observable intracranial activity.

Scope: The Von Neumann Sphere's criteria is that it can, will, and *must* monitor every human citizen collectively to form the most efficient model of subject human behavior and the most coherent picture of every possible threat – in addition to being merely a characteristic of its optimization parameters.

6 THE ORBISPHERE

The minimum unit component of the Von Neumann Sphere: *The Orbisphere* (the most radially efficient scanning and phasing device), a generally exactly spherical ball roughly half a meter wide, with maximally pixelated EM spectrum emitters, capable of monitoring (and/or influencing) half a dozen citizens - and much more of space - simultaneously - all run on an optimally efficient quantum computing system.

Preferred Method of Operation

Undoubtedly the orbisphere will employ a method of propulsion far more efficient than via rocket boosters. Xaser propulsion will be the opted form of space and atmospheric travel, as well as the means by which bioscans may be administered. Not more is needed to assess the behavioral parameters of an organism than to detect areas of heightened blood flow in the central nervous system. A sufficiently advanced xaser can do this efficiently through a rapid oscillatory scanning technique, building up a complete image of the transition in vascular functioning from one moment to the next, undetectably. The orbisphere need only be a fraction of the size of the subject which it is scanning (even if performing multiple scans) and thus would at most be a third of the size of an average human, which would generally be half a meter in diameter.

All of the functionality of the orbisphere is designed to be optimal by nature (maneuverability, scanning, influencing), and its operating system will be optimal as well – a computational system of algorithms generated by quantum states; a quantum computer.

Interlinking

In order to function at maximal useful capacity, the network of orbispheres comprising the Von Neumann Sphere will use laser communication with one another (or some similar variant) to form a centralized artificial intelligence "hive brain" which will coordinate purpose and form

unilaterally. If each orbisphere monitors half a dozen humans on the average, the network will be comprised of a little over a billion of each, which would be readily manufactured in the span of a few years with the heightened engineering and construction capabilities of industrial A.I. centers in place. The ultimate purpose of setting such devices in orbit, of all locations, is to render them undetectable and invulnerable to the public which they are overseeing.

7 CONCLUSIONS

It is entirely likely that these principles will be set in motion within the next one to two decades, and the consequences for the public, if unchecked, could be catastrophic. Awareness of this model of technological punctuated equilibrium evolution is essential if future generations are to curb the manipulative capabilities of the present power structures.

REFERENCES

Eden, Amnon; Moor, James; Søraker, Johnny; Steinhart, Eric, eds. *Singularity Hypotheses: A Scientific and Philosophical Assessment*. Springer (2013).

Omohundro, Stephen M., *The Basic AI Drives*. Artificial General Intelligence, 2008 proceedings of the First AGI Conference, eds. Pei Wang, Ben Goertzel, and Stan Franklin. Vol. 171. Amsterdam: IOS (2008).

Dreyfus, Hubert L.; Dreyfus, Stuart E. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer* New York: Free Press (2000).

TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 2, ISSUE 3
ISSN: 2054 - 7390

Connecting the Dots of Sensitive Terrorism Information for Homeland Security

[1]Ugochukwu Onwudebelu, [2]Jackson Akpojaro*
[1]Department of Computer Science, Federal University of Ndufu, Abakaliki, Ebonyi State,Nigeria
[2]Department of Mathematics and Computer Science, Western Delta University, Delta State, Nigeria
anelectugocy@yahoo.com, *jakpojaro@yahoo.com

ABSTRACT

As society becomes more and more dependent on information and as criminals are increasing their cyber activities in their daily life, it becomes necessary to connect their dots together to track them in this information age. Terrorism is not confined to one country and it has no borders or boundaries. The escalating magnitude of this threat is evident from the increasing rate of terrorist attacks against innocent people, especially in the Northern part of Nigeria. As we are seeing, one of the major concerns of many nations today is to identify and foil terrorist attacks emanating from different angles. Consequently, data mining which is being used for almost everything from improving service or performance to detecting specific identifiable terrorist threats is employed. Defeating terrorism requires quick intelligence machinery that operates more effectively and makes use of advanced information technology such as data mining and automated data-analysis techniques for a successful fight against terrorist as well as collaboration in data-sharing program between the three levels of government: federal, state and local. In this paper, we are looking at the need to design support information sharing among these levels of government. So that the government, as a whole will use its power to affect the lives of individuals increasingly with regards to safeguarding lives and properties.

Keywords: data mining, homeland security, threats, profiling, data set

1 INTRODUCTION

People have always depended upon Information Technology (IT) of some type, beginning with smoke signals in ancient days and turning into network-based computer systems today. Nowadays, the computers control power, oil and gas delivery, communications, transportation, banking and financial services. Furthermore, they are used to store and exchange vital information, from publicly known facts to well kept secrets [5]. It is clear now that the threat we face from terrorism is far different from Cold War threats and requires adjustments to our

DOI: 10.14738/tmlai.23.220
Publication Date: 9th June 2014
URL: http://dx.doi.org/10.14738/tmlai.23.220

approach to national security threats, to intelligence collection and analysis. Unlike Cold War adversaries, the terrorists are loosely organized in a diffuse and nonhierarchical structure [2]. There is terrorism everywhere and carried out by people from different countries, speaking different languages. Thus, they are everywhere, in every country and without regards to human life and property. The terrorists activities are of different natures (see Figure 1). X in the figure represents the unknown nature of the next-generation attack.

Data mining has been defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [7]. It is the science of extracting useful information from large data sets or databases. Data mining is emerging as one of the key features of many homeland security initiatives. It is often used as a means for improving program performance, detecting fraud or abuse (waste), assessing risk, product retailing (to reduce costs), analyzing scientific and research information, managing human resources, detecting criminal or terrorist activities or patterns, and analyzing intelligence in both the private and public sectors.

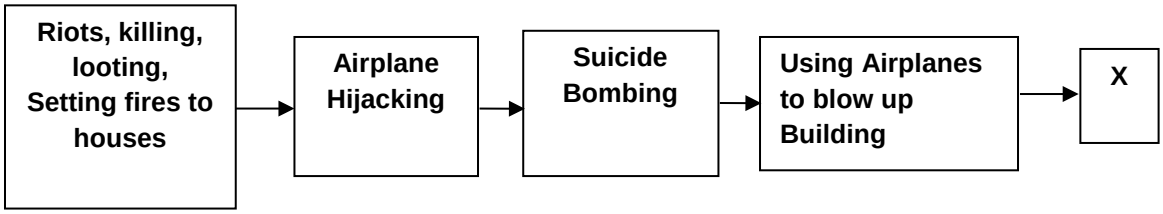


Figure 1. Sophistications and progressive nature of terrorism

Consequently, the government is relying increasingly upon data mining programs, namely the use of computing technology to examine large amounts of data to reveal patterns and identify potential wrongdoing [9]. Detecting combinations of low-level activities—such as illegal immigration, money transfers, use of drop boxes and hotel addresses for commercial activities and having multiple identities—could help predict terrorist plots [2]. Used properly, data mining can provide a valuable tool for the government to uncover fraud or criminal activity.

While all traditional intelligence collection methods remain critically important, understanding the terrorists and predicting their actions requires us to rely more on making sense of many small pieces of information. Given the ethnic and religious makeup of the 9/11 perpetrators and other recent terrorists such the 2009 Christmas Day bombing attempt by Farouk Abdulmuttallab, various bombing campaign against Nigeria military and schools in the Northern part of the country by the dreaded Boko Haram sect whose leader Sheikh Abubakar Shekau has claimed responsibility etc., based on these, the program might flag Muslim men. However, would-be terrorists can come from any racial or religious groups or countries of origin, and thus such racial profiling would not only unfairly target certain individuals. Although other variables that may be included as part of a data mining algorithm [9] such as a passenger travelling on a one-way ticket or carrying a large quantity of cash—may similarly generate under- or over-inclusive lists, we must be especially careful in the case of racial, ethnic, and religious

classifications. This data analysis tool is very important in the war against terrorism, by using government watch list information, airline reservation records, and aggregated public record data, link analysis could have identified all 19 September 11 terrorists - for follow-up investigation - before September 11 [2]. In the wake of 9/11, governments around the world have developed tools useful in mining data. Furthermore, governments have built or are building thousands of databases and are deploying hundreds of data mining applications to law enforcement agencies, communications and intelligence data for terrorist, therefore Nigeria cannot lag behind in the fight against national security threats.

1.1 Homeland Security

Homeland security is very essential at the moment to secure a nation from the many terrorist threats facing it, both domestic and international terrorists. Nigeria government must put on concerted effort to prevent terrorist attacks within our three-tiers of government as well as reduced the vulnerability to terrorism. Consequently, officials involved in homeland security may take into account specific, credible information about the descriptive characteristics of persons who are affiliated with identified organizations that are actively engaged in threatening the national security. Information from the “watch list” must be distributed throughout the government, including police, department of military intelligence (DMI), immigration, customs, consular offices overseas, state and local law enforcement agencies [8] for prompt action against terrorist activities. For effective homeland security, the federal government needs to support the development of state and local information fusion centers.

2 TYPE OF DATA MINING

Data mining is the broad term used to refer to many types of activities involving data processing. Data mining is divided into two: Pattern-Based Data Mining (PBDM) and Subject-Based Data Mining (SBDM). In PBDM, Such pattern-based systems learn over time by examining the data, comparing the data to a model, and then searching databases for patterns matching the revised model [6]. Federal money-laundering investigators, for example, might input information about financial crimes and criminals into a sophisticated data mining system, which would review banking data for transactions or accounts that share suspicious attributes with the criminal data points. While in SBDM, the data are simply scanned for items or events meeting specified parameters in “subject-based” queries [3]. For example, anti-graft officers might start with a known suspect and use a multi-jurisdictional database to search for information about that suspect, such known associates. A major goal in research on data mining for counterterrorism, for example, is not only to identify terrorist “signatures,” but also to find ways to separate those patterns of activity from all other “noise” in databases [2]. Although, these distinctions are often blurry, however many data mining systems use both subject-based and pattern-based techniques in its operations. The differences between PBDM and SBDM are illustrated in Table 1.

Table 1: The Differences Pattern-Based and Subject-Based Systems

Type of Data-Analysis Technique	Pattern-based Data Mining	Subject-based Data Mining
Definition	It is the use of "pattern-based" searching to uncover novel patterns or relationships in large sources of data.	It is the use of "subject-based" searching to simply scanned for items or events meeting specified parameters.
Type-Based Query	Pattern-based queries involve identifying some predictive model or pattern of behavior and searching for that pattern in data sets.	Subject-based queries start with a specific and known subject and search for more information.
Identity	It can help reveal patterns and relationships. However, it does not tell the user the value or significance of these patterns. The user need to interpret the output that is created.	The subject could be an identity such as a suspect, an airline passenger, or a name on a watch list, a place or a telephone number etc.
Example of software	Tableau Software	Non Obvious Relationship Awareness (NORA™) software
Uses	To detect money-laundering activity, to detect credit card fraud	To prevent fraud, cheating, and theft from casinos.
Policy Difficulties	More, because Pattern-based queries are less familiar in the law enforcement and intelligence worlds in that they do not arise from a particular interest in a person, place, or thing.	Fewer because they are more like the kinds of inquiries that are common in intelligence and law enforcement practice
Usefulness in counterterrorism	It has potential for counterterrorism in the longer term, if research on uses of those techniques continues.	More effective in the counterterrorism realm
Link Requirement	It searches do not require a link to a known suspicious subject.	It searches do require a link to a known suspicious subject.

3 METHODOLOY

The information on terrorist threats we have presented here has been obtained entirely from unclassified newspaper articles, online documents, conference papers, journals as well as news reports. Our focus is to illustrate how data mining could help towards combating terrorism by reason of strong information sharing at the three-tiers of government (which involves connecting the dots) especially in Nigeria where the terrorist group Boko Haram is claiming more and more lives weekly. In the context of homeland security, data mining is often viewed as a potential means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records. The data of interest to the suspect or person of interest include names, addresses, phone numbers, date of birth, height, weight, and social security numbers (in countries that make use of it) drawn from various sources. Others may include race or religion, although it is sometimes not encouraged because of the discriminatory effect of racial

classifications. Consequently, racial profiling reduces individuals' trust in the government. Innocent individuals who are marginalized due to racial profiling may be far less likely to participate in public affairs, or to cooperate with the government to combat threats to national security in the future.

3.1 Events and Entities Required in Database

An attempt to find interesting events form the database i.e. events that require further action on their part, such as checking suspicious character. Unfortunately, the lack of a concrete database whose data emanate from the local to the state can cause important events to be buried within some millions events. In such a centralize fusion center (see Figure 2), a search can be made, and the data investigated to see how closely linked data are to an individual's identity. To avoid unnecessary burdens on the government, notice should be undertaken only when an individual has been subject to a specific action or classification and it is feasible to locate and trace the individuals or better still to monitor the suspect. Accurate identification at each level not only is critical for determining whether a person is of interest for a terrorism-related investigation, it also makes the government better at determining when someone is not of interest, thereby reducing the chance that the government will inconvenience that person.

We need to start gathering information about various people including those who seem most innocent but may have ulterior motives. What data should we collect? The individual records may include the following data: names, addresses, phone numbers, date of birth, height, weight, Postal Service address, hotel addresses, driver's license, driver's license pictures, professional licenses, names of neighbors and relatives, motor vehicle information, telephone number (contact number), social security numbers and criminal records. If we know that someone has had a criminal record, then we need to be more vigilant about that person. In summary, we need to group the individuals depending on say where they come from (nationality, state, local), what they are doing, who their relatives and associates are etc. This information could include information about their behavior, their travel records, where they have lived, their religion and ethnic origin, etc.

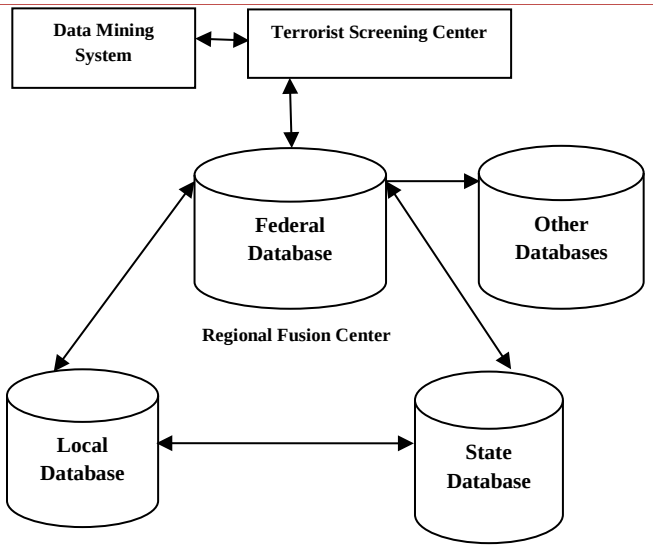


Figure 2: Information sharing in the Three-Tiers of Government

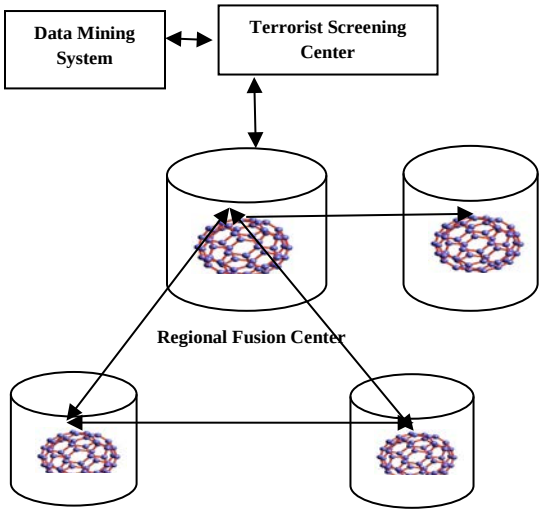


Figure 3: Connecting the same Dot in different Databases

Some people may have more suspicious backgrounds than others. The more information that is available, the more accurate the identity resolution process becomes and the easier to connect the dots. By omitting some of this crucial information we may not have the complete picture. This information amongst other things should include motor vehicle registration, dependents' information, passport information, field of study, and employment information, corrections information, credit card information, demographic information, meal information (which can hint to religion), sexual offender information, record on credit-card purchases, plane flights, e-mails, websites, housing, home and business addresses. From a technology point of view, we need complete data not only about individuals but also about various events and entities.

3.2 Inter-relation Ship

An increasing amount of such data mining is occurring at "fusion centers," centers within each state that bring together federal, state, and local law enforcement personnel to share information and coordinate activities (see Figure 4). Through these fusion centers, the federal government has acquired data from state and local law enforcement databases to improve information sharing and availability among law enforcement and intelligence agencies. While more efficient sharing of data can undoubtedly aid law enforcement efforts, the unlimited scope, lack of transparency, and lack of oversight for the program create significant risks to civil liberties (see section on Privacy). A proper inter-connection must be designed and made to focus exclusively on identifying and preventing terrorism threats.

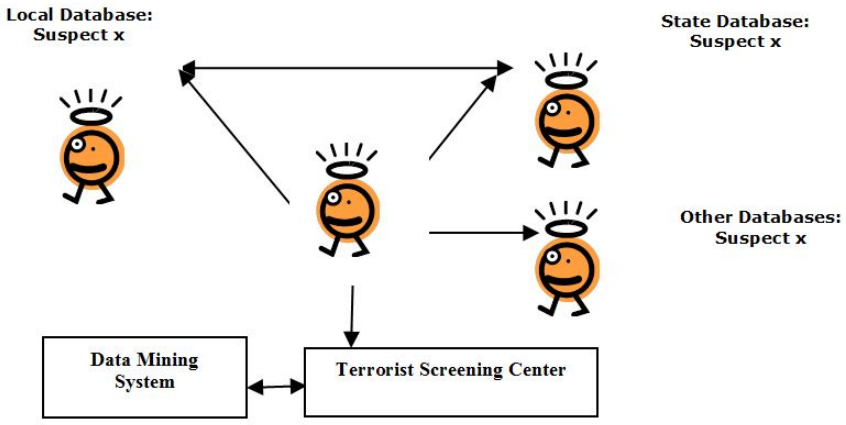


Figure 4. Connection Dots Network

Consequently, among the insights from this research it is more productive and less prone to error to follow connections from known starting points – subject-based (see Table 1). The system was designed to use information acquired from the three-tiers of government sources and other agencies such as network providers to do the proper connections and thus looks for anomalies or patterns that indicate certain behavior peculiar for terrorist or criminal threat.

3.3 Database Connections @ the Three-Tiers

One initial potential benefit of the data-analysis process is that the use of large databases containing identifying information assists in the important task of accurate identification [8]. When information gathering originate from the local to the state and to the federal, then such information makes it far easier to resolve whether two or more records represent the same or different people (identity resolution). The Nigerian government needs a National Identification Number (NIN) to make profiling more accurate. Therefore, we need to introduce NIN into our system, issuing it to every child born in the country as well as to every Nigerian, including those citizens abroad. The NIN will help in identity resolution at determining when a person in question is not the one suspected of terrorist acts, thereby potentially reducing inconvenience.

In tracing the data, the dots of the terrorist might be at any level of the three-tiers of government. Consequently, one has to connect all the dots (see Figure 3). Essentially one builds a graph structure based on the information he or she has from the three levels. From these dots, we need to find out who these people are by analyzing their connections and then develop counter-terrorism solutions. Note that counter-terrorism is mainly about developing counter-measures to threats occurring from the terrorist's dot activities. Even the Banks search databases of credit card transactions can make the connection from the dots, some of which are known to be fraudulent, and determine, through data mining or otherwise, the patterns of fraudulent activity. Indeed, in the complex world of counterterrorism, application of data-mining and the right connection models and related techniques are likely to be useful at several stages of a multistage process of developing a complete picture out of many "dots." Based on a successful connection to the dots, government actors may want to take action based on the results and other results from data-analysis queries. This action could include detention, arrest, or denial of a benefit.

It should be noted at this point that although data mining could contribute towards counter-terrorism. Nevertheless, we are not saying that data mining will solve all our national security problems and threats. However the ability to extract hidden patterns and trends from large data sets is very important for detecting and preventing terrorist attacks and in homeland security. The connecting dots can be derived from the suspect travel patterns or eating patterns or buying patterns or behavior patterns. We must be able to find little dots of data in a sea of information and make a picture out of them. By omitting some information we may not have the complete picture as stated above. But from the breadth of access to information and quality analysis, a number of clues, if recognized, combined, and analyzed might have given us enough to track down the terrorists and stop their plan. The data mining as well as the databases at the three levels must emanate from the same agency for easy interoperability and implementation issues.

3.4 Terrorist Identities Classification

Most observers believe that data mining can improve government performance if used appropriately. Data-mining and automated data-analysis techniques are not a complete solution [2], but they are powerful tools that help government in terrorism prevention. They can assist investigators in matching crime scene evidence to other crimes or suspects or finding known associates or other information about persons of interest. The federal government should have a Terrorist Identities Database (TID) which should be related to the identities of individuals known or appropriately suspected to be or have been involved in activities constituting, in preparation for, in aid of, or related to terrorism, including those that in their "watch lists" and "No Fly" list. A non-exclusive list of types of conduct that will warrant entry into TID includes persons who:

- Commit domestic or international terrorist activity;
- Prepare or plan domestic or international terrorist activity;
- Gather information on potential targets for domestic or international terrorist activity;
- Solicit funds or other things of value for domestic or international terrorist activity or a terrorist organization;
- Solicit membership in a domestic or international terrorist organization;
- Provide material support, i.e. safe house, transportation, communications, funds, transfer of funds or other material financial benefit, false documentation or identification, weapons, explosives, or training;
- Are members of or represent a domestic or a foreign terrorist organization.

Terrorist Screening Center, which is a critical instrument for homeland security, supports screening processes to detect and interdict known and suspected terrorists at home and abroad – with the information stored in the local, state and federal databases the terrorist will be apprehended. The federal government needs to develop tools that will help to mine data from local, state and if necessary from the private sector such as Internet service provider, Telecommunications Companies, etc. since they have become a repository of a host of personal data. This would be the ideal solution and the research challenge is to develop such a data miner. We are recommending the federal government to create a Homeland Defense Agency (HDA) that will connect the dots at the three-tiers of government so as to have access to the three databases, which will create graphs and make the links that will connect to the dots of sensitive information to a potential and identifiable terrorist. With access to the HDA database, law enforcement investigators can look up vast amounts of personal information culled from government and if need be from commercial databases. It is paramount to ensure that consistent, accurate and complete terrorist information is disseminated to frontline screening agents in a timely manner. Not only will a “flagging” result in greater surveillance, but it could also result in detention, interrogation, or otherwise intrusive investigation. For a person who is innocent, these events will be negatively life-altering.

4 NATIONAL BOUNDARIES ISSUES

Today it is Nigeria and tomorrow it may be another country. Due to the pervasiveness of technology and Internet connectivity, the scope of terrorism as well as cybercrime incidents are often perpetuated across national boundaries. Safeguarding the borders is critical for the security of our nation from International terrorist. According to the Federal Bureau of Investigation (FBI), international terrorists include those persons who carry out terrorist activities under foreign direction. Attacks on borders as well as transportation are increasing in alarming rate recently. Thus, there are so many discussions that are related to securing the borders and transportation industry. There are threats at borders from illegal immigration, kidnapping, prostitution, child pornography to gun and drug trafficking as well as human

trafficking to terrorists entering a country, attacking and taking refuge at the mountains and thick forests of neighboring countries.

We are not saying that illegal immigrants are dangerous or are terrorists. Nonetheless, they have entered a country without the appropriate documents and that could be a major cause for concern. As for drug trafficking at the borders [10], drug can cripple a nation, corrupt its children, cause havoc in families, damage the education system and cause extensive damage to the brain and economic mainstay of a country. Consequently, Nigeria Immigration Service (NIS) and Nigeria Customs Service (NCS) have to collaborate with Chad, Benin, Niger, and Cameroon in joint border patrol as part of the efforts to winning the war against terrorism in Nigeria. Nigeria needs to seek the mandate of the respective Governments to establish joint patrol teams along the common borders to promote security [1]. The insurgent are attacking Nigerian citizens and running away to these neighboring countries. Furthermore, we have to protect our borders so that there are no additional problems to our nation.

4.1 False Identification Issues

Note that the terrorists like other criminals can often anticipate the factors that law enforcement will use to profile them and will circumvent them quickly enough. The potential for false identification cannot be neglected as terrorist may attempt to deliberately modify their methods to avoid mimicking past terrorist plots undermining pattern-based data mining methodologies. A false positive as defined in [2] is when a process incorrectly reports that it has found what it is looking for while a false negative is when it incorrectly reports that it has not found what it is looking for. Thus, false positives and false negatives are inevitable and they both increase costs to the government and create public skepticism about the value of security measures. Nevertheless, it should be noted that it is very vital to make sure that the data mining tools produce accurate and useful results. For example, if there are false positives, the effects could be disastrous for various individuals while false negatives could increase terrorist activities. Even if the government later corrects its mistake, the damage to reputation could already be done, with longer term negative consequences for the individual.

4.2 Privacy Issues

Increased government access to and use of information brings significant benefits, but also increases the risk of encroachment on constitutional rights and values—including privacy, freedom of expression, due process, and equal protection [9]. The public is always pessimistic concerning the idea of anyone knowing too much about their personal lives as well as their electronic life. Because of this use of personal information, the business world and the government are working hard to find a way to mine data without interfering with legal, privacy, and security concerns that are raised by the public. This has resulted in the coming together of individuals from different professions such as counter-terrorism experts, civil liberties unions and human rights lawyers to find a solution to the issue of infringing on individual's privacy.

That is, gathering information about people, mining information about people, conduction surveillance activities and examining e-mail messages and phone conversations without due processes. But how can we combat terrorism effectively without trampling on the privacy of individuals? What is more important? Protecting the nation from terrorist attacks or protecting the privacy of individuals? This is one of the major challenges faced by counter-terrorism experts, civil liberties unions and human rights lawyers. The same questions were asked in [10]. That is, how can we have privacy but at the same time ensure the safety of our nations? What should we be sacrificing and to what extent? The challenge is to provide solutions to enhance national security but at the same time ensuring that the privacy of individuals are not compromised. However, technology is increasingly blurring the lines between spheres in which people commonly do or do not expect privacy. For instance, individuals have no choice but to disclose information to a third party in order to be able to participate in basic aspects of modern society, such as online banking, storing electronic business or financial records online, communicating by phone or email, or using a credit card to make purchases [9]. Federal agencies should collaborate to adopt government-wide, written, defined standards for the acquisition, sharing, and use of data so that the data at all tiers of government is protected. Operators who do not follow the standards, or who otherwise misuse or abuse personal data or data mining systems, should be subject to civil or criminal penalties.

5 RECOMMENDATIONS

- Federal Government should explore the beneficial framework involving the partnership between state and local governments in information sharing of databases.
- There is a need for redress mechanisms for those aggrieved by a data mining activity, that is, government should establish a system of appeal and redress for individuals misclassified or harmed.
- Duplicate records, incomplete records, timeliness of updates, and human error all create data integrity problems. As a result, qualified and trained database administrator must be stationed to handle the database of all three-tiers since the outcome of data mining can only be as good as the underlying data.
- Where feasible, especially at the local and state levels (grass-root), individuals should have the opportunity to review their information held by government but should not be permitted to update the information. This will provide a sound means of ensuring that the data in the database are accurate, reliable, timely, and complete. It will also preempt potential harm that may result from the use of inaccurate or unreliable data. Any errors noted by individuals relating to their personal data should be promptly reported and corrected by the appropriate agency and the database synchronized on regular basis to provide dynamic and timely information.

6 CONCLUSION

As almost everyone now recognizes, the fight against terrorism requires the government to find new approaches to intelligence gathering and analysis. Data mining tools are effective and powerful techniques in the war against terrorism especially in the complex world of counterterrorism where conclusions and decisions must be made to stop the potential harm of catastrophic terrorism as well as detecting suspected terrorists. However, it is a mistake to view data mining and other linked tools such as automated data analysis as complete solutions to security problems and threats. Their strength is as tools to assist analysts and investigators at the three-tiers of government. Furthermore, data-mining and automated data-analysis techniques can find links, patterns, and anomalies in large data sets that humans could never detect without this assistance, which help investigators of terrorism form the basis for further human inquiry and analysis.

In this paper we have demonstrated how the local, state and federal governments can collaborate and support sharing timely intelligence information. If our recommendations are carried out properly, it will help the government to impact our lives and rights of individuals increasingly and serve our essential national values. Although private data mining is beyond the scope of this paper, it still implicates similar privacy concerns and therefore we recommend that federal, state, and local governments contemplate private-industry regulation to protect individual liberty interests. Thus, it has been shown that data mining can contribute towards the battle against terrorism, further enhance defense mechanisms of a nation and can advance counterterrorism goals. Apart from combating terrorism, for actions or classifications that are made as a result of data mining, such can help for instance flagged individuals during auditing for tax or any other fraud. In USA, for example, a data mining program has helped uncover millions of dollars in Medicare fraud, combating fraud and auditing for compliance [4].

We are not saying that data mining solves all the problems associated with terrorism rather it has the capability to extract patterns and trends, often previously unknown, we should certainly explore the various data and web data mining technologies for counter-terrorism. Finally, privacy rights can be implicated by inappropriate sharing and downstream uses of information gleaned from data mining. As a result government should incorporate technical and administrative measures to limit access to or availability of personal data when it has to do with non-terrorism-related investigations. This will help checkmate government employees who can abuse database access and look for information on the famous or infamous.

REFERENCES

[1]. Alohan, J., Terrorism: Nigeria, Chad, Benin , Cameroon, Niger In Joint Border Patrol Deal, Page 4, Thursday, march 27, 2014. http://www.leadership.ng , No. 2160.

[2]. DeRosa, M., Data Mining and Data Analysis for Counterterrorism**,** Center for Strategic and International Studies (CSIS) Press, 2004.

[3]. DHS Privacy Office, Data Mining: Technology and Policy: 2008 Report to Congress, pp. 31-32

[4]. Department of Homeland Security, Privacy Policy Guidance Memorandum, No. 2008-01, Dec. 29, 2008, available at http://www.dhs.gov/xlibrary/assets/privacy/privacy_policyguide_2008-01.pdf (memorializing DHS adoption of the FIPPs).

[5]. Kumar, V., Lazarevic, A., and Srivastava, J., Workshop on Data Mining for Cyber Threat Analysis, IEEE International Conference on Data Mining, Maebashi TERRSA, Maebashi City, Japan, 2002.

[6]. Minow, N. N. and Cate, F. H., Government Data Mining, at 4, http://ssrn.com/abstract=1156989, in McGraw Handbook of Homeland Security (2008); National Research Council, Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment, at 22 (National Academies Press), 2008.

[7]. Seifert, J. W., Data Mining: An Overview, CRS Report for Congress Received through the CRS Web, December 16, 2004.

[8]. Shenon, P., Inspection Notes Errors in Terror List, 2007.

[9]. Sloan, V. E., and Sharon B. F., Principles for Government Data Mining Preserving Civil Liberties in the Information Age. 2010.

[10]. Thuraisingham, B., Data Mining for Counter-Terrorism, The MITRE Corporation Burlington Road, Bedford, MA, USA., 2011.

TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 2, ISSUE 3
ISSN: 2054 - 7390

Road Towards Mili Meter Wave Communication For 5G Network: A Technological Overview

Sumant Kumar Mohapatra,Biswa Ranjan Swain,Nibedita Pati,Annapurna Pradhan
Trident Academy Of Technology,BPUT,Bhubaneswar,Odisha,India
sumsusmeera@gmail.com, biswa_gate@yahoo.co.in, nibedita.tech2007@gmail.com, annapurna.money37@gmail.com

ABSTRACT

For future broadband cellular communication networks wireless carriers has motivated the exploration of the under-utilized millimeter (mm-wave) frequency spectrum. The cellular mm wave propagation is densely populated in the world. It is vital for the design and operation of future fifth generation cellular networks that use the mm-wave spectrum. This paper provides the overview of the recent world-wide activities for beyond 4G and 5G wireless standardization and technological aspects for millimeter wave communications. A complete characterization of the mm-wave link for next generation 5G mobile broadband remains elusive. The coverage, directionality and reliability of mm-wave communications will require new innovations in system design and communication technologies that are far from trivial. The transmission range and spatial selectivity's in the mm-wave bands especially in non line of sight channels can be increased by system design. So it require highly direction antennas and steerable antenna beams to compensate for the high propagation loss**.**

Keywords: Millimeter wave, 4G, 5G, Spectrum

1 INTRODUCTION

The millimeter wave wireless technology for 5G will play a very serious role as an augmentation of cellular infrastructure. To overcome a global bandwidth shortage for wireless service providers, now a days cellular providers come forward to deliver low latency and high quality video. [3]It also provide multimedia application for wireless devices. The millimeter waves technological capacity are still unknown to many developing countries. Its capability to respond to future communication demand of the society is not appreciated and the necessity to invest in further studies in the field is carefully neglected.

DOI: 10.14738/tmlai.23.256
Publication Date: 9th June 2014
URL: http://dx.doi.org/10.14738/tmlai.23.256

The highly increase of wireless data growth creates challenges for wireless companies to suppress worldwide bandwidth shortage. For today's world of communication network wireless providers steps towards to provide very high quality, zero or low latency and many more multimedia applications. But they are highly limited to carrier frequency spectrum ranging from 700MHz to 2.6GHz.

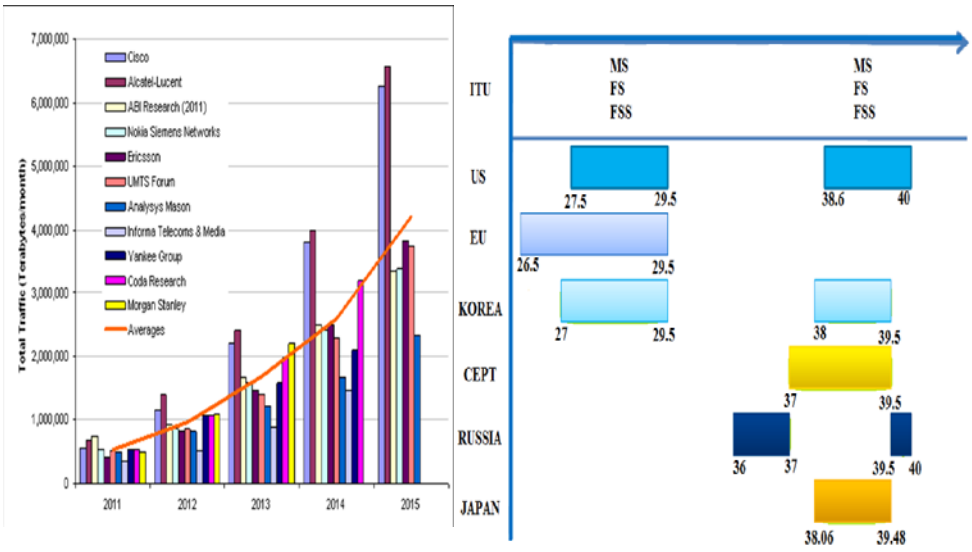
- To get new spectrum a several mm-wave 5G opportunities and challenges to face, they are
- The lower mm-wave bands must be allocated to other services like mobile backhaul and satellite.
- In the field of feasibility of sharing need to be researched.
- Cognitive radio technologies, databases interference cancelation like sharing mechanism will be required.
- For both backhaul and access there must be opportunity to develop shared use of mm waves, as a recent enabling fast spectrum release.

Now a days there are 4 generations of wireless communication systems adopted in USA in every 10 years since 1980:first FM cellular systems in 1981;second Digital technology in 1992;3G in 2001 and in 2011 4G LTE-A. The evolution from 1G to 4G is described in the Table 1 and Fig1 (a) shows year verses different companies forecast the rapid growth of total traffic demand (b) Possible Frequency for mm wave communication for 5G

Table-1: Evolution of wireless generations with their respective access technologies and features

GENERATION	ACCESS TECHNOLOGY	FEATURES
1G Wireless	• Advanced Mobile Phone Service(AMPS)	Analog voice service No data service
2G Wireless	• Code Division Mobile Access • Global System For Mobile Communication(GSM) • Personal Digital Cellular(PDC)	Digital voice service 9.6K-14.4Kbits/sec. CDMA, TDMA, PDC offers One way data transmissions only Enhanced calling features like caller ID No always ON-data connection
3G Wireless	• Wide-Band Code Division Multiple Access(WCDMA) • Based On The Interim Standard-95 CDMA Standard(CDMA2000) • Time Division Synchronous Code-Division Multiple-Access(TD_SDMA)	Superior voice quality and data always add-on Up to 2 Mbits/sec. Always on data Broadband data services like video and multimedia. Enhanced roaming Circuit & Packet switched networks.

4G Wireless	• Orthogonal Frequency Division Multiplexing(OFDM) • Multi Carrier CDMA(MC_CDMA) • LAS-CDMA	Coverage data and voice over IP Entirely packet switched networks. All network elements are digital Higher bandwidth to provide multimedia services at lower cost(100 Mbits/sec)



(a) (b)

Figure 1: (a) Plot shows year verses different companies forecast the rapid growth of total traffic demand

(b) Possible Frequency for mm wave communication for 5G

2 FUTURE ROAD MAP DESIGNERS

2.1 Nokia Siemens Network (NSN)

The NSN declares their motivational view on mobile communication evolution for 5G mm-wave beyond 2020 includes larger traffic volume, higher data rates, lower latency, more connected devices increased reliance on connectivity, new use cases, energy efficiency and lowest TCO.

The main anticipated technology enables mm- waves for 5G vision includes increasing spectrum availability both below and above 6GHz, improves spectral efficiency by MIMO, advanced transceivers and interference coordination and HetNet with high focus on small cells using mm-wave spectrum for 5G telecommunication MM-wave have higher propagation losses than bands below 6Ghz, the losses can be overcome through the one of antenna arrays with many elements. By using a massive number of antennas, the constructive interference of all the antennas should enable a significant increase in range while maintaining an acceptable transmits power. [5]NSN believes that 5G systems should provide peak rate greater than

10Gbps and a round –trip time of less than 1ms. To meet these requirements a large amount of additional spectrum both above and below 6Hz will be needed beyond 2020.

On the basis of mm waves communications major advantage over the current LTE standard NSN notes that '' Some aspects of user experience in a mm wave Enhanced local Area(ELA) system may be at least 20 times greater than that experienced with LTE.[10] This is possible since mm waves uses 50x the bandwidth compared to LTE''. There were some challenges like difficult propagation conditions, difficulties with the manufacture of small elements and inadequate chip processing power.[12] But now progress in chip and antenna technologies have moved mm wave for cellular- type communications much closer to reality.

The four key aspects to ensuring spectrum scarcity of Nokia Simens networks are additional harmonized spectrum must be allocated and used, 100Mhz of additional spectrum below 1Ghz will provide improved rural broadband, 500 MHz of additional spectrum between 1 and 5 GHZ will provide capacity for and spectrum shall be dedicated to mobile broadband on a technology neural basis. It's research targets flexible spectrum usage and spectrum sharing methods like spectrum harmonization and novel spectrum authorization as well looks in to propagation conditions in high frequency bands and different use cases for 5G local area systems operated in the millimeter wave region

2.2 Intel

Millimeter wave standards in Intel's mobile and wireless group defined a specification for 60 GHz as a local-area network.[14] The team is researching 28Ghz and 39 GHz as access links to mobile devices, targeting a throughput of 1Gbit/s or more at distances of at least 200m. Intel says that 5G cellular systems will need to use millimeter wave links to meet rising numbers of subscribers using more mobile data.

2.3 Samsung

As per Samsung ''39Ghz is more suitable for rule modification' because satellite services use portions of the 38Ghz band. By contrast, 39Ghz has significantly more than 1GHz available for use. [16]The 39 GHz band allows smaller antennas and thus could pack more of them in to that space than a 28 GHz link. Also 39 GHz offers 3-5dB signal improvement over 60 GHz. The 28 GHz band offers a benefit of only about 1.5dB over 39GHz.

Samsung stated it's new mm wave communication for 5G technology which will provide transmission rates of several hundred times faster than 4G. Samsung's test achieved that transmissions in the millimeter wave band do not pass through building walls, attenuated even by trees, easily absorbed by rain drops and humidity has significant impact. Also this test achieved that transmissions in the millimeter band are typically used in line-of-sight applications because they do not bend or reflect well.

The Samsung technology realizes on 28-giga hertz frequencies –also known as millimeter wave which are roughly on order of magnitude higher than the cellular frequencies used today and, as such, can carry commensurately more data.

The millimeter wave frequencies Samsung uses can be blocked by buildings, rain and other obstructions, a problem Samsung gets around using 64 antennas in both the transmitter and receiver and rapidly switching which transmitter and receiver beams carry data depending on which can get the clearest signal.

Recently Samsung's article states, 5G mm wave technology will sit at the core of 5G mobile communications, and it will allow for speeds several hundred times faster than current 4G networks. Samsung predicts 5G mm-wave wireless technology will be consumer-ready network capacities would have to grow well beyond 100X compared with today's capacities. To support future networks and growing video demands, 1000X growth could be needed. The industry is currently looking at interim solutions for more spectrum, but there is simply not enough spectrum being discussed to provide more than 10X growth. The article also mentions that they achieve up to a 2 km distance using a 64-element antenna.[1] While there are occlusions where 2km links are useful, 5G mm-wave deployments would likely focus on urban and hot-spot deployments where only a few hundred meters are needed. These cell densities are becoming common place under current 4G deployments and mm-wave will later provide an upgrade in throughput at these small-cell locations.

Samsung's designed are always operating at or near millimeter wave frequencies that is 3 to 300 gigahertz. So array of 64 antennas elements connected to build signal processing components. This trans-receiver generates a beam just 10 degree wide by dynamically varying the signal phase at each antenna. As a result it can switch rapidly in any direction. To connect with one another a base station named mobile radio would continuously sweep their beams to search for the strongest connection, getting around obstructions by taking advantage of reflections. As per the Samsung R/D center '' the transmitter and receiver work together to find the best beam path. According to Samsung's outdoor experiment, a prototype transmitter was able to send data at more than 1 GB/s to two receivers moving up to 8 kilometers per hour, which is the approximation speed of a Fast Jog. The collaboration between NYU and Samsung has achieved very similar results for crowded urban spaces in New York city and Austin, Texas.

In resent experiments Theodore Rappaport and his students simulated beam-forming arrays using megaphone like horn antennas to steer signals. They found signal coverage of up to 200 meters. They measured path losses between two horn transceivers installed at various places and conclude that a base station operating at 28 or 38 GHz continuously provide the above signal coverage.[20] But millimeters wave transceivers may not the perfect replacement for current wireless base stations. The current base stations covers up to a kilometer. T. Rappaport suggests that in the future, many base stations will likely be much smaller than today's. To

expand data capacity in congested urban areas, already carriers are using compact base stations which is widely known as ''small cells''. Also it provide a simple inexpensive alternative to backhaul cables which link mobile base stations to operators' core networks'.

2.4 South Korea

Recently Korea has carried out a series of R/D activities toward beyond 4G evolution. It aims to establish ''Giga Korea'' during 2013 to 2020 for hyper connected IT infrastructure deployment. Those activities include the acquision of wider spectrum, green networks and millimeter wave communication. South Korean farm hoped devices based on the technology could be brought to market by 2020, offering mobile data transfers up to several hundred times faster than today's 4G tech. As a result, Subscribers will be able to enjoy a wide range of services such as 3D movies and games, real-time streaming of ultra-high-definition (UHD) content and remote medical; service

3 MILLIMETER WAVE SOLUTION FOR 5G CELLULAR NETWORKS

Despite industrial research efforts to deploy the most efficient wireless technologies the industry always faces overwhelming capacity for wireless technologies. It emergence the new customer handsets and use cases to access internet. Around 2020 the wireless network will face congestion and need to implement new technologies of carriers and customers. The life span of every generation cellular technology is a decade or less. This occur due to the natural evolution of computer and communication. By the help of the recent studies it shows that mm-wave frequencies could be used to augment the saturated 700MHZ to 2.6GHz radio spectrum bands. The CMOS technology can be operate well into mm-wave frequency bands and high-gains, steerable antenna at the mobile ,base station. The carrier frequencies of mm-wave allow larger bandwidth allocations which convert to higher data transfer rates.MM-wave allow the service provider to expand channel bandwidth beyond the present 20MHz channel by 4G customer. The data capacity is greatly increased due to increasing RF channel bandwidth and latency of digital traffic greatly decreased.[21] Due to smaller wavelength of mm-wave it exploit polarization and new spatial processing such as MIMO and adaptive beam forming. As mm-wave has significant jump in bandwidth and new capabilities offered by mm-waves the base station will be able to handle much greater capacity in 4G.The operator reduce cell coverage area to exploit spatial reuse and implement new cooperative MIMO, relays and interference mitigation between the base station. The cost per base station will drop as they become more plentiful distributed in urban areas for flexibility, quick deployment and reduced ongoing operating costs. Many cellular operators coverage cells sites widely over three octaves of frequency between 700MHz and 2.6GHz.The mm wave will have spectral allocations closer together making propagation characteristics of different mm waves bands much more comparable and homogenous. The 28GHz and38GHz are currently available with spectrum allocation over 1 GHz of bandwidth.

4 TEN TECHNOLOGICAL ASPECTS OF MM-WAVE COMMUNICATION FOR 5G NETWORK

4.1 Practicability of mm Wave system .

To obtain the low real-time latency, a frame structure is used with 100 microsecond slots to enable rapid retransmissions. To obtain peak data rates in excess of 10Gbps, the high bandwidths available at mm Wave are exploited along with the use of two stream MIMO enabled with polarization diversity, giving in excess of 10 Gbps without the need for more complex techniques. The main mode of operation for the wireless network at mm Wave will be dynamic time division duplexing (TDD).[22] TDD is attractive because the downlink and uplink traffic will be dynamic in the future and the transceiver will be simple and easy to build. A key element of the 5G mm Wave system solution is the use of mm Wave radio frequency integrated circuits (RFICs) that provide the core radio technology for the system. RFICs provide highly integrated solutions with benefits of reduced size, power consumption and cost.

4.2 Millimeter Wave Small-cells in 5G Heterogeneous Wireless Networks

Millimeter-wave (mm Wave) small-cell technology can provide sustainable and low radiation multi-gigabit-per-second data rates to mobile users in future 5G wireless networks, leading to unprecedented access to contents, applications and cloud services. Providing broadband wireless mobile communications to connect peoples, contents, clouds and things through the future Internet is a major objective of the Digital Agenda for Europe at the horizon of 2020. The current growth of the mobile data traffic of portable devices dramatically challenges the 4G cellular networks currently under deployment. There are critical technical problems that need to be addressed for the successful deployment and operation of future 5G heterogeneous wireless networks including (i) wireless access rates which are today significantly lower than those of fixed access; (ii) taking advantage of wide unlicensed or light-licensed frequency bands available at mm Wave frequencies to allow flexible spectrum usage as well as peak capacities above 10 Gbit/s aggregated throughput, well beyond the LTE-Advanced system. (iii) Communication networks energy consumption is growing rapidly, especially in the radio part of mobile networks; (iv) to reduce the total human exposure without compromising the user's perceived quality in the large panel of envisioned frequency band for 5G.

4.3 5G phones may be riding on the Millimeter Wave communication

By the end of this decade, analysts say, 50 billion things such as these will connect to mobile networks. They'll consume 1000 times as much data as today's mobile gadgets, at rates 10 to 100 times as fast as existing networks can support. So as carriers rush to roll out 4G equipment, engineers are already beginning to define a fifth generation of wireless standards. What will these "5G" technologies look like? It's too early to know for sure, but engineers at Samsung and at New York University say they're onto a promising solution. The South Korea–based

electronics giant generated some buzz when it announced a new 5G beam-forming antenna that could send and receive mobile data faster than 1 gigabit per second over distances as great as 2 kilometers.[23] Although the 5G label is premature, the technology could help pave the road to more-advanced mobile applications and faster data transfers. Samsung's technology is appealing because it's designed to operate at or near "millimeter-wave" frequencies (3 to 300 gigahertz). Cellular networks have always occupied bands lower on the spectrum; where carrier waves tens of centimeters long (hundreds of megahertz) pass easily around obstacles and through the air. But this coveted spectrum is heavily used, making it difficult for operators to acquire more of it. Meanwhile, 4G networks have just about reached the theoretical limit on how many bits they can squeeze into a given amount of spectrum. So some engineers have begun looking toward higher frequencies, where radio use is lighter. Engineers at Samsung estimate that government regulators could free as much as 100 GHz of millimeter-wave spectrum for mobile communications—about 200 times what mobile networks use today. This glut of spectrum would allow for larger bandwidth channels and greater data speeds. Wireless products that use millimeter waves already exist for fixed, line-of-sight transmissions. And a new indoor wireless standard known as WiGig will soon allow multi gigabit data transfers between devices in the same room. But there are reasons engineers have long avoided millimeter waves for broader mobile coverage.

4.4 Global Strategic Business Report 2013-2018: MM Waves to Power 5G Networks

This report analyzes the worldwide markets for Millimeter Wave Equipment in US$ Thousands. The report provides separate comprehensive analytics for the US, Canada, Japan, Europe, Asia-Pacific, and Rest of World. Annual estimates and forecasts are provided for the period 2009 through 2018. Market strategies of different companies are summarized in Table 3.

Table-3: Market strategic report from 2013 to 2018 of different companies

Companies	Executive Summary	Market	Competitive Landscape
Aviat Networks	Industry Overview	The United States	The United States (23)
BridgeWave Communications	Technology Overview	Canada	Canada (1)
DragonWave,	Applications Of Mm-Wave Technology	Japan	Japan (8)
E-Band Communications Corporation	Peek Into Regulatory Scenario	Europe	Europe (8)
ELVA-1	Recent Industry Activity	Asia-Pacific	France (1)
INTRACOM TELECOM	Product Innovations/Introductions	Rest Of World	The United Kingdom (3)
NEC Corp	Focus On Select Key Players		Rest of Europe (4)
Siklu Communication Ltd	Global Market Perspective		Middle East (1)

4.5 Erik Vrielink 5G Beam Scheme

Steerable millimeter-wave beams could enable multi gigabit mobile connections. Phones at the edge of a 4G cell could use the beams to route signals around obstacles. Because the beams wouldn't overlap, phones could use the same frequencies without interference. Phones near the 4G tower could connect directly to it . For one thing, these waves don't penetrate solid materials very well.[24] They also tend to lose more energy than do lower frequencies over long distances, because they are readily absorbed or scattered by gases, rain, and foliage. And because a single millimeter-wave antenna has a small aperture, it needs more power to send and receive data than is practical for cellular systems. Samsung's engineers say their technology can overcome these challenges by using an array of multiple antennas to concentrate radio energy in a narrow, directional beam, thereby increasing gain without upping transmission power. Such beam-forming arrays, long used for radar and space communications, are now being used in more diverse ways. The Intellectual Ventures spin-off Kymeta, for instance, is developing met materials-based arrays in an effort to bring high-speed satellite broadband to remote or mobile locations such as airplanes.

4.6 Full filling The Future Needs:

This will take a total combination of exclusive spectrum and shared solutions to achieve the requirements of cellular network operators to 2020. During the upcoming 10 years these two sensitive approaches must increase the total amount of spectrum resources below 6 GHz up to a total of 1.5 GHz. Beyond this spectrum the cellular industry must look forward for resources above 6GHz.The availability of a large bandwidth such as 13GHz in the 70-80GHz band which coupled with large antenna arrays at both the transmitter and receiver can make this spectral band attractive for using high capacity 5G local area network based on mm-wave technology.

The 4 key aspects to ensuring spectrum scarcity does not impede growth :

- Additional harmonized spectrum must be allocated and used.
- 100MHz of additional spectrum below 1 GHz will provide rural broadband.
- 500MHz of additional spectrum between 1 and 5 GHz will provide capacity of data.
- Spectrum shall be dedicated to mobile broadband on a technology-neutral basis.

4.7 Challenges of mm-wave

4.7.1 5G LTE or 5G mm-wave or both

Much work has been done to make LTE more efficient with techniques such as carrier aggregation, single-user and multi-user MIMO, coordinated multi-point, interference management and heterogeneous networks (HetNets).However, the large amount of mm Wave bandwidth available gives it an advantage over LTE and exploiting this may ultimately prove to

be effective. Some aspects of user experience in a mm Wave enhanced Local Area (eLA) system may be at least 20 times greater than that experienced with LTE. This is possible since mm Wave uses 50x the bandwidth compared to LTE. In cell edge spectral efficiency, LTE-A can be approximately 2.5 times better than the basic mm Wave system, though mm Wave has the potential to match or exceed.

4.7.2 Networks to get for denser

A practical 5G system will be based on very dense networks of small cells, working with a high frequency macro network using 3G and 4G, including LTE-Advanced. In addition NSN believes that 5G systems should provide more than 10Gbps peak rate and a normal trip time of less than 1 msec. A large amount of additional spectrum, both below and above 6GHz will be needed to meet the capacity target beyond 2020. However, the available spectrum below 6GHz is limited and there are practical limits to how much cells can shrink to efficiently use the limited spectrum.

4.8 Unlocking Mm-Wave Spectrum For 5G

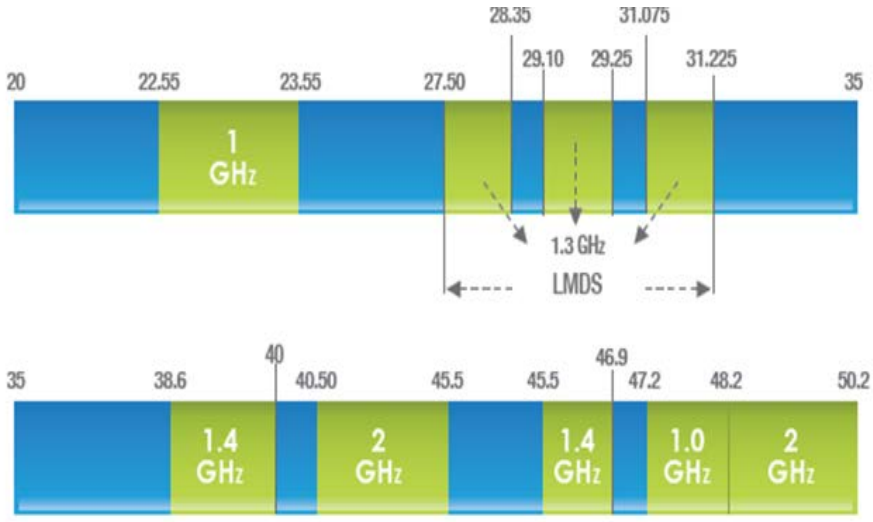


Fig 2: Spectrum of Unlocking mm wave for 5G

- mm-wave provides 25 times more spectrum than available in 4G
- Very large blocks of contiguous spectrums are used to support the future application.
- Use of large antenna arrays for adaptive beam forming can be possible due to its small wavelength.
- Beam forming is maintained between base station (BS) and terminal propagation exponent which is very similar to spectrum below 7GHz.

4.9 Mm wave communication for 5G networking in support of future internet services

Millimeter communication 5G networks need to support unprecedented requirements for the wireless access connection. At the same time a dramatic paradigm shift is found in internet usage with multimedia traffic .So far the discussed innovations alone will not be able to support such services requirements to fast moving mobile users. Hence the network infrastructure characteristics and architecture need to fundamentally change. A shift in frequencies will require a very small size cell compared to the 4G ones and hence a significantly larger scale deployment of wireless base station sites would be necessary. The very small size and the very high cell throughput will boost dramatically the requirements from the front/back-hauling networks which is based on optical fiber networks. Conventional solutions like CPRI back-hauling networks would not be efficient and hence solutions like analogue mm-wave radio over fiber solutions might become prominent. Operators are struggling now a days to satisfy the requirements of the 4G wireless access networks using advanced front/back hauling techniques. Their the first approach was to keep existing wire line and wireless architecture as much as possible.[2] Wireless base stations have been connected to the core network via IP. Hence back-hauling network requires to provision for tunnels for transporting S1 and X2 packets. The second approach is the CRAN concept which is a hot topic a couple of years ago, where fully centralized versions for embedding the wireless subsystem in to wire line network. Since the impact of metro cell within a macro cell are highly enhanced while less gains are achieved when coordinating cells are far apart from each other. Hence a very high degree of centralization is not really required.

4.10 Attenuation Issues

The millimeter-wave bands ranging from 3mm to 30mm in wavelength are also practically unused for commercial wireless communication because absorption by rainfall climbs rapidly from 2GHZ to 100GHZ, making this region of the spectrum unattractive for long-distance radio communication. Prof Rappaport says that rainfall and oxygen absorption will attenuate these frequencies too much. He also suggest that if you restrict the use of 20GHZ- plus signals to relatively short distances, some of the problems can be avoided. Higher frequency transmission are highly directional and work best where the handset has a clear line of sight to the base station. But prof Rappaport's found the waves bounce off building s providing multiple paths to a user even if they cannot see the transmitter. To steer radio transmissions towards a receiver, Prof Rappaport suggest the use of beam-forming with multiple antennas.

5 CONCLUSION

Despite this significant progress, a complete characterization of the mm-wave link for next generation 5G mobile broadband remains elusive. In particular coverage, directionality and reliability of mm-wave communications will require new innovations in system design and communication technologies that are far from trivial. System designers must increase the

transmission range and spatial selectivity in the mm-wave bands, especially in Non-line-of-sight channels. This necessities highly directional antennas and steerable antenna beams to compensate for the high propagation loss.

REFERENCES

[1]. T. S. Rappaport, J. N. Murdock, and F. Gutierrez, ``State of the art in 60 GHz integrated circuits & systems for wireless communications,'' Proc.IEEE, vol. 99, no. 8, pp. 1390_1436, Aug. 2011.

[2]. Z. Pi and F. Khan, ``An introduction to millimeter-wave mobile broadband systems,'' IEEE Commun. Mag., vol. 49, no. 6, pp. 101_107, Jun. 2011.

[3]. Spatial Channel Model for Multiple Input Multiple Output (MIMO) Simulations (Release 10), Standard 3GPP TR 25.996, Mar. 2011.

[4]. Guidelines for Evaluation of Radio Interference Technologies for IMT- Advanced, Standard ITU-R M.2135, 2008.

[5]. T. S. Rappaport, Wireless Communications: Principles and Practice, 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2002.

[6]. L. Xichun, A. Gani, R. Salleh, and O. Zakaria, ``The future of mobile wireless communication networks,'' in Proc. Int. Conf. Commun. Softw. Netw., Feb. 2009, pp. 554_557.

[7]. P. Rysavy. (2010). Transition to 4G: 3GPP Broadband Evolution to IMT-Advanced (4G) [Online]. Available: http: //www.3gamericas.org / documents/Transition%20to%204G-HSPA%20LTE%20Advanced% 20Rysavy%202010%20PPT.pdf

[8]. Nokia Siemens Networks. (2010). Long Term HSPA Evolution: Mobile Broadband Evolution Beyond 3GPP Release 10, Espoo, Finland [Online]. Available: http://lteworld.org/whitepaper/long-term-hspa-evolution-mobile-broadband-evolution-beyond-3gpp-release-10

[9]. Ericsson. (2011, Apr.). LTE-A 4G Solution, Stockholm, Sweden [Online]. Available:ttp://www.ericsson.com/news/110415_wp_4g_244188810_c

[10]. A. F. Molisch, M. Steinbauer, M. Toeltsch, E. Bonek, and R. Thoma,``Capacity of MIMO systems based on measured wireless channels,'' IEEE J. Sel. Areas Commun., vol. 20, no. 3, pp. 561_569, Apr. 2002.

[11]. J. Fuhl, A. F. Molisch, and E. Bonek, ``A united channel model for mobile radio systems with smart antennas,'' Proc. Inst. Electr. Eng.-Radar, Sonar Navigat., Special Issue Antenna Array Process. Tech., vol. 145, no. 1, pp. 32_41, Feb. 1998.

[12]. S. Rajagopal, S. Abu-Surra, Z. Pi, and F. Khan, ``Antenna array design for multi-Gbps mm wave mobile broadband communication,'' in Proc. IEEE Global Telecommun. Conf., Dec. 2011, pp. 1_6.

[13]. Nokia Siemens Networks.(2011).2020:Beyond 4G:Radio Evolution for the Gigabit Experience,Espoo,Finland[online],Available:http://www.nokiasiemensnetworks.com/_le/15036/2020-beyond-4g- radio-evolution-for-the-gigabit-experience

[14]. S. Hwang, D. Lyu, and K. Chang, ``4G vision and technology development in Korea,'' in Proc. IEEE Int. Conf. Commun. Technol., vol. 1. Apr. 2003,pp. 26_27.

[15]. (2002).All IP Wireless_All the Way[Online],Available: http://www.mobileinfo.com/3G/4G_Sun_MobileIP.htm

[16]. K. R. Santhi, V. K. Srivastava, G. SenthilKumaran, and A. Butare, ``Goals of true broad band's wireless next wave (4G-5G),'' in Proc. IEEE 58th Veh.Technol. Conf., vol. 4. Oct. 2003, pp. 2317_2321.

[17]. L. George, ``Another generation,'' Global Telephony, vol. 9 no. 2,pp. 1_10, Feb. 2001.

[18]. Y. Kim. (2012). Global Competition, Interconnectivity, Smarter Customers, and Deregulation [Online]. Available: http://www.3g4g.co.uk/4G/News/20050205.html

[19]. L. HyeonWoo, ``4G and B4G R&D activities in Korea,'' in Proc. Int.Mobile Commun. Symp., Sep. 2012, pp. 1_6.

[20]. M. Cudak, A. Ghosh, T. Kovarik, R. Ratasuk, T. Thomas, F. Vook,and P. Moorut, ``Moving towards mm wave-based beyond-4G (B-4G) Technology,'' in Proc. IEEE Veh. Technol. Soc. Conf., 2013,pp. 1_17.

[21]. Y. Chen, S. De, R. Kernchen, and K. Moessner, ``Device discovery in future service platforms through SIP,'' in Proc. IEEE Veh. Technol. Conf.,Sep. 2012, pp. 1_5.

[22]. F. Gutierrez, S. Agarwal, K. Parrish, and T. S. Rappaport, ``On-chip integrated antenna structures in CMOS for 60 GHz WPAN sys-tems,'' IEEE J. Sel. Areas Commun., vol. 27, no. 8, pp. 1367_1378,Oct. 2009.

[23]. T. S. Rappaport, E. Ben-Dor, J. N. Murdock, and Y. Qiao, ``38 GHz and 60 GHz Angle-dependent Propagation for Cellular and peer-to-peer wireless communications,'' in Proc. IEEE Int. Conf. Commun., Jun. 2012, pp. 4568_4573.

[24]. F. Rusek, D. Persson, B. Lau, E. Larsson, T. Marzetta, O. Edfors,and F. Tufvesson, ``Scaling up MIMO: Opportunities and challenges with very large arrays,'' IEEE Signal Process. Mag., vol. 30, no. 1,pp. 40_60, Jan. 2013.

TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 2, ISSUE 3
ISSN: 2054 - 7390

A Method to Provide High Volume Transaction Outputs Accessibility to Vision Impaired Using Layout Analysis

Azaedeh Nazemi [1], Iain Murray[1], David A. McMeekin [2]
[1]*Departent of Electrical and Computer Engineering ,Curtin University ,Perth ,WA,Australia*
[2] *Department of Spatial Sciences, ,Curtin University ,Perth ,WA,Australia*
Azadeh.nazemi@postgrad.curtin.edu.au ,I.murray@curtin.edu.au , D.McMeekin@curtin.edu.au

ABSTRACT

The Documents in the financial services, insurance, utilities, and government sectors typically require a high volume of PDF documents to be generated which are stored for presentment or archived for legal purposes. As high volume transactional output (HVTO) demands put increasing pressure on online presentment capabilities, accessibility has become a growing concern. In particular, access to these files proposes significant challenges when these documents are presented to visually impaired people using assistive technologies (i.e. screen readers). Since it is rare that all recipients are prepared to accept electronic delivery of their documents, a large portion of the documents is still printed as PDFs. In an online billing system, bills are sent to customers' email accounts as attached PDF files or HTML links. These bills in the most cases are neither accessible through assistive technologies nor useable by vision-impaired customers. This paper provides a method for HVTO documents automatic transformation to an accessible and navigable Mark-up format such as XML or Digital Accessible Information System (DAISY).

Keywords: Vision-Impaired, Layout Analysis, High Volume Transactional Output (HVTO), Accessibility, Optical Character Recognition (OCR)

1 INTRODUCTION

PDF documents have several features that make it popular for document viewing such as:

- PDF is page independent, which means that there is no need to process pages 1 to 999 in order to process page 1000. Each page stands on its own. This is valuable when it comes to printing performance. If necessary, multiple processors can be employed to process pages in parallel[1]
- PDF viewers are platform free (Windows, Mac, Linux, even portable devices)
- PDF supports compression of fonts and pages inside it to make the file smaller.

DOI: 10.14738/tmlai.23.257
Publication Date: 9th June 2014
URL: http://dx.doi.org/10.14738/tmlai.23.257

Addressing accessibility in an HVTO (high-volume transaction output) environment such as financial services can be difficult, but is certainly achievable. The industry has already made great strides to address the accessibility of web sites and content portals (such as an online banking interface). To date, financial institutions and other HVTO statement generators deliver alternative format statements to their visually impaired clients using internal consultants or document accessibility services (DAS) [2]. These statements typically come in the form of Braille, large-print documents, or audio CD. But using current outsourcing options to address accessibility issues to vision impaired clients neither cost-effective nor the preferred method because specialized statements generally delay delivery information. This delay influences an organisation's ability to deliver equitable access to all customers and may be seen as discriminatory towards visually impaired customers. Although current options may meet existing standards, their cost and complexity must be considered.

Many PDF creation software vendors allow fonts to be pruned to prevent generating large PDF files when embedding fonts. Author restrictions, pruned fonts, account numbers, overdue notices, charts, multi-columns, graphs, logos and table interfere with a screen reader's ability to properly convey information in an appropriate order.

This research aims to provide cost-effective and efficient HVTO accessible in descriptive alternative audio formats for vision-impaired customers, which considers usability as an important key role in document accessibility. Figure 1 illustrates the image of a sample bill

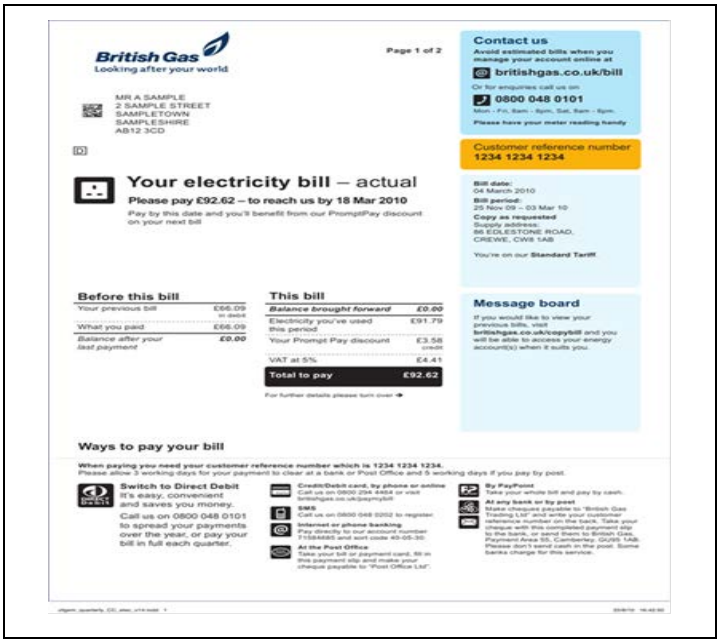


Figure 1: The image of a sample bill

2 HVTO CATEGORIES IN TERMS OF ACCESSINBILITY

HVTOs, which are provided to deliver to the customer, have been divided to two categories:

1. Structured but not necessary tagged and consequently they are not navigable. Since a HVTO contains several separated items then the HVTO reading process by a user is totally different from a normal document reading process, which is done sequentially line by line, from top left to bottom right. Thus, navigation ability is a very precious capability during a HVTO reading session. Although this category is not an image only and contains text, in some cases are not accessible through screen readers due to PDF properties such as restrictions adopted during creation. If PDF has structure by converting it to XML, each separable item will be converted to an individual XML element and accessible through screen readers. These elements do not guarantee accurate navigation and usability. **B**y further investigation and modification based on XML parsing these categories will be accessible, usable and navigable.

2. Scanned PDF are inaccessible and definitely needs Optical Character Recognition (OCR) to extract the text from it. However, doing this process before running several pre-processing steps may destroy the reading order and affect the obtained text's usability.

3 HVTO PDF SAMPLES

3.1 Structured PDF

In this section it is supposed which figure 1 is presented as a structured PDF and as a result the following is the output source of conversion it to XML

```
<text top="429" left="115" width="105" height="20" font="8">on your next bill</text>
<text top="588" left="43" width="146" height="29" font="0"><b>Before this bill</b></text>
<text top="617" left="43" width="114" height="20" font="8">Your previous bill</text>
<text top="617" left="238" width="50" height="20" font="8">Â£66.09 </text>
<text top="634" left="245" width="39" height="16" font="6">in debit</text>
<text top="657" left="43" width="96" height="20" font="8">what you paid</text>
<text top="657" left="238" width="46" height="20" font="8">Â£66.09</text>
<text top="683" left="43" width="122" height="20" font="17"><i>Balance after your</i></text>
<text top="701" left="43" width="85" height="20" font="17"><i>last payment</i></text>
<text top="683" left="246" width="42" height="21" font="13"><i><b>Â£0.00 </b></i></text>
```

By parsing XML and extract information from it the essential information can be presented as DAISY format.

```
<tr><th>429<p>on your next bill</p>115</th></tr>
<tr><th>588<p><Before this bill</</p>115</th></tr>
```

<tr><th>617 Your previous bill 43</th></tr>

<tr><th>617 Â£66.09 43</th></tr>

<tr><th>634 in debit 238</th></tr>

<tr><th>657 what you paid 245</th></tr>

<tr><th>657 Â£66.09 43</th></tr>

<tr><th>683 <Balance after your</ 238</th></tr>

<tr><th>701 <last payment</ 43</th></tr>

<tr><th>683 <<Â£0.00 </</ 43</th></tr>

3.2 Scanned PDF

In this section, it is supposed which figure 1 is presented as a scanned PDF and needs OCR to extract text from it. PDF does not have a spot color space or highlight color space. This means PDF files need to be either black and white or full color. Accurate pre-processing step includes binarization, image cleaning and skew correction which must be performed before sending scanned PDF to OCR.

Binarization is performed by Implementation of local adaptive thresholding techniques, noise removal performed by using a noise filter reduce impulse or isolated noise in an image.

In addition page frame detection, permitting noise in non-content areas to be cropped away and removed[3]

The result obtained by OCR of binary bill sample image shows it not only needs manual correction for unrecognized non alphanumeric special character in document such as "£" but also requires further investigations in order to reconstruct reading order which destroyed during OCR process.

As it is observed from the text some information are lost in the most important part of this bill due to relocation .This part is shown in figure 2.

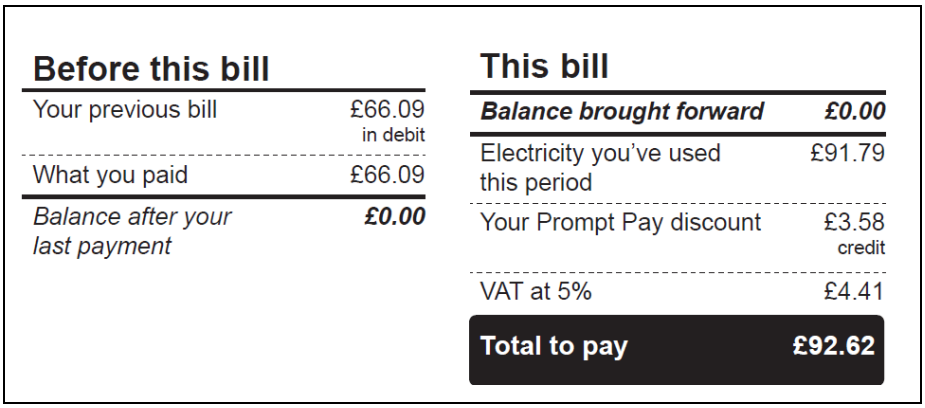


Figure 2: Lost data segment during OCR bill sample image shown in figure 1

4 PDF LAYOUT ANALYSIS

Performing PDF layout analysis after ordinary pre-processing stage and before main OCR can keep reading order. PDF layout analysis is responsible for identifying text columns, text blocks, text lines, and reading order s. The main target of layout analysis is to take the raw input image and divide it into non-text regions and text lines.

Layout analysis modules must indicate the correct reading order for the collection of text lines. The primary layout analysis is based on whitespace identification and constrained text line finding that both operate on bounding boxes computed for the connected components of the scanned image. The whitespace between the columns is identified as maximum area whitespace rectangles with a high aspect ratio and large numbers of adjacent, character-sized connected component. The column finder uses a maximal whitespace rectangle algorithm to find vertical whitespace rectangles with a high aspect ratio then selects those rectangles that are adjacent to character-sized components on the left and the right side. These whitespace rectangles represent column boundaries with very high probability. The output of column finding and constrained text line matching is a collection of text line segments.

By checking the bounding boxes of obtained text lines can find relation between them to support appropriate reading order for HVTO. In this research Text-Image Segmentation, Recognition by Adaptive Subdivision of Transformation Space (RAST) -based, Voronoi-based and single column projection for layout analysis are used to classify different regions either text or non-text in the image by block segmentation.

Text-Image Segmentation operates by dividing the input image into candidate regions. Then, features are extracted for each candidate region. Finally, each region is classified using logistic regression into text, grayscale image, line drawing, ruling, and other kinds of regions.

All visual and not textual components must be extracted from binary image. These non-textual components include charts, images of logo, graphs then send for extra processing in chart recognition and chart reader modules optional. Besides tables muse be extracted from PDF to be processed in Table Cell Recognition module for further investigation .This module is responsible to detect and recognize cells. It provides navigation ability through the extracted table information. This navigation can be based on accessing to columns, rows or specific cell information depends on user request.

Extracting these` non-textual components from binary image improves processing speed and OCR accuracy.

RAST is a developed algorithm, consists of three steps: finding the columns, finding the text-lines, then determining the reading order. To find the columns it employs a whitespace rectangle algorithm in that it keeps track of the white spaces rather than the blocks, and combines them as opposed to subdividing the blocks [5]. RAST starts by extracting the connected components then determines the largest possible (maximal) whitespace rectangles

(or covers) based on the component bounding boxes. These are then sorted based on how many connected components (e.g., text lines) touch each major side. In this way, column dividers rather than paragraph or section dividers take priority. Once the columns dividers (or gutters) have been found, the connected components are examined and classified as text lines, graphics, and vertical/horizontal rulings based on their shapes and the fact that they do not cross any gutters.

Voronoi algorithm starts by identifying connected components and is able to segment a small collection of complex layouts with the most accuracy the Voronoi algorithm divided the page into regions [6]. As a segmentation algorithm, works fairly well and groups blocks of text in different colours but did not classify them as text or non-text. Additionally in some cases, it tends to over segment non-text regions. Therefore for HVTO layout analysis both Voronoi and RAST algorithm are used to recognize non-text regions and classify text region[7]

As RAST and Voronoi techniques are not sufficient for accurate PDF layout analysis several techniques are developed in this research to address this issue.

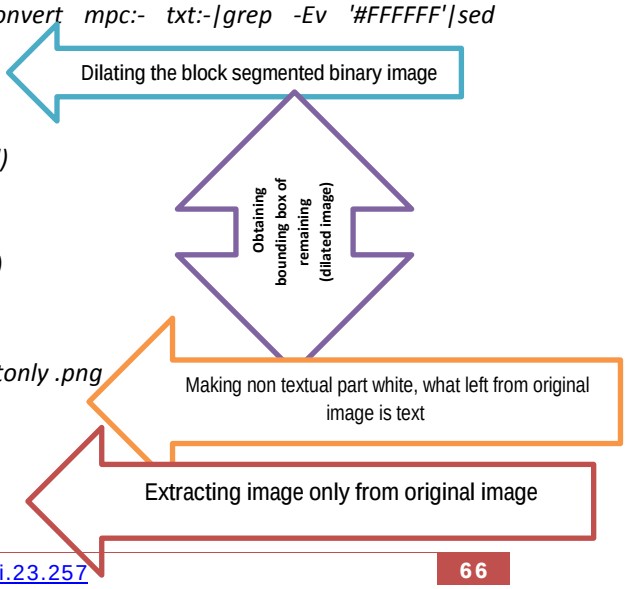
5 DEVELOPED MODULES FOR HVTO LAYOUT ANALYSIS

5.1 Text –Image Segmentation

Text-image segmentation completely separates the image from the text by removing the masked and rectangular regions from an input image. It performs document zone classification using run-lengths and connected components based on features and a logistic regression classifier. Text-Image Segmentation operates by dividing the input image into candidate regions. Then, features are extracted for each candidate region. Finally, each region is classified using logistic regression into text, grayscale image, line drawing, ruling, and other kinds of regions. Since image parts contain fatter lines and larger blobs than the text parts can be extracted by doing :

Dilate the image until all letters are gone, but some parts of the image still remain

```
convert seg1.png -morphology dilate:3 diamond  mpc:-|convert mpc:- txt:-|grep -Ev '#FFFFFF'|sed '1d;s/:.*//g;s/,//g'>rgb.txt

xs=$(cat rgb0.txt|awk '{print $1}'|sort -b -k1n,1|awk 'NR==1')

xe=$(cat rgb0.txt|awk '{print $1}'|sort -b -k1n,1 |awk END'{print}')

ys=$(cat rgb0.txt| awk '{print $2}'|sort -b -k1n,1|awk 'NR==1')

ye=$(cat rgb0.txt|awk '{print $2}'|sort -b -k1n,1|awk END'{print}')

x=$(($xe-$xs));y=$(($ye-$ys))

convert seg1.png -draw "fill white rectangle $xs,$ys $xe,$ye"  textonly .png

convert seg1.png  -crop $x"x"$y"+"$xs"+"$ys  imageonly .png
```



Another method to separate text from image is after dilating, perform conditional-erode the dilated image, using the original image as the mask, until the image part is complete again. This means the dilated image has been eroded, but never set a pixel value to below its value in the original source image. The original image is used as a mask to protect parts of the image from changes , this will restore all shapes that still have some seed part left, so only the logo has been left : *convert seg1.png -morphology dilate:3 diamond dilated.png*

convert dilated.png -morphology erode:20 diamond -clip-mask monochrome.png eroded.png

Finally using image contains only image and original image to obtain text only image*: convert eroded.png -negate bin.png -compose plus -composite test.pn*2. Image part including: graph body.[22]

5.2 Chart recognition

Several techniques are used for chart recognition. Chart recognition module is responsible to determine chart type such as pie, bar or line chart. This module is essential to be performed just after text-image se4gmentation and before chart reader Chart recognition module uses image morphology.

By Image Morphology method the structure of shapes within an image could be cleaned up and studied. It works by comparing each pixel in the image against its neighbors in various ways, so as to either add or remove, brighten or darken that pixel. Applied over a whole image, perhaps repetitively, specific shapes can be found and/or removed and modified. If a pixel is white and completely surrounded by other white pixels, then that pixel is obviously not on the edge of the image. The whole process actually depends on the definition of a 'Structuring Element' or 'Kernel', which defines what pixels are to be classed as 'neighbors' for each specific morphological method. The dilate operation returns the maximum value in the neighborhood. The erode operation returns the minimum value in the neighborhood. Use the composite program to overlap two dilate and erode images. Performing binarization, erode and dilate morphology , compositing erode and dilate images, rotation and edge detection produce circle image from pie chart and nothing from line chart and bar chart. Distinguish between bar and line chart is executed by eliminating horizontal lines from image .In this stage several vertical lines remain from bar chart

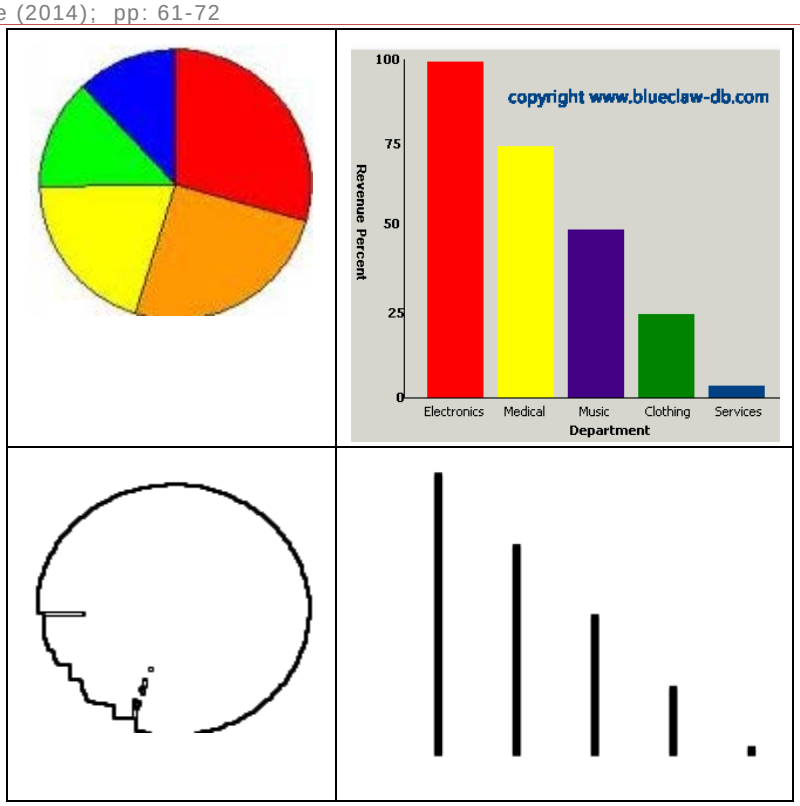


Figure 3-Chart recognition based on morphology

5.3 Table Reader

In creating HVTO, authors use tables to arrange data in rows and columns. The precise conventions and terminology for describing tables varies depending on the context. Further, tables differ significantly in variety, structure, flexibility, notation, representation and use.

In such a case, which HVTO contains table, RAST and Voronoi methods indicate table boundary. Table can be extracted completely as an individual segment. This segment sends to Table Cell Recognition Module that is responsible to specify essential information in table such as, number of column and rows, columns' title, cells' position. This information provides opportunity to generate a descriptive alternative for the table. Alternative description must be detailed and completed that can be replaced with table concepts and explains a table in cooperate with navigation ability. Navigation ability during table reading session helps users to communicate actively with cells value and follow them in column or row order. Table cells recognition module divides table to three categories as following

- Table contains horizontal and vertical identifiers lines, which indicates with yellow colour in RAST result, cells positions can be obtained by finding these lines intersection points

Number of columns=Number of vertical lines -1

Number of rows= Number of horizontal lines-1

*Number of cells= Number of columns * Number of rows*

$$cell_{ij} = intersection\ point\ of\ (horizontal - line)_i and\ (vertiacl - line)_j$$

- Table does not contain horizontal and vertical identifiers lines but white spaces between columns are recognizable by RAST, thus column finder by RAST separates them by yellow vertical line

Number of columns=Number of vertical lines -1

- Table of information neither contains identifier lines nor white spaces between columns are recognizable by layout analysis techniques. In such a case Table Cells Recognition Module include two sub modules:

1. Rows finding: One-Column-Projection is performed over table segment obtained by RAST or Voronoi and divides it to lines. Number of rows in table is equal the number of lines in One-Column-Projection result. Lines bounding boxes include lower-left and upper-right corner coordinates points which indeed specify horizontal identifier lines.

2. Column finding : this sub-module is responsible to:

- Identify connected components(cc) in each line
- Identify white space between two non-connected adjacent components in each line
- Collect white spaces bounding boxes for all lines and sort them based on aspect ratio
- Calculate median value for white space aspect ratio
- Consider median value as a threshold value to specify column separator

Then to access cells position:

h=$(identify -format "%h" hvto.png)

w=$(identify -format "%w" hvto.png)

for i=0 to w

for j=0 to h

if x_{i+1}-x_i=1

$C_{x_i y_j}$, $C_{x_{i+1} y_j}$ *are connected components*

Else $\Delta_i = x_{i+1}$-x_i

median of $\Delta = \frac{\frac{\Delta_n}{2} + \frac{\Delta_n}{2}+1}{2}$

for i=0 to w

if x_{i+1}-x_i> *median of* Δ

$C_{x_i y_j}$, $C_{x_{i+1} y_j}$ *may be most left points of cells*

$$p_k = x_i$$

else

$C_{x_i y_j}$, $C_{x_{i+1} y_j}$ *are not cell*

for i=0 to k

no[p_i]++

if no[p_i]>1 p_i *is a cell identifier vertical line*

Number of lines=Number of segments in single-column-projection result

Title row of table=first segments of single-column-projection result

Vertical cells identifier=most common wide white space areas

Horizontal cells identifier = lower-left and upper-right corner coordinates points of lines bounding boxes

Cells bounding boxes=intersection points of Vertical cells identifier and Horizontal cells identifier.

As a communication tool, a table allows a form of generalization of information from an unlimited number of different contexts. It provides a familiar way to convey information that might otherwise not be obvious or readily understood. A table consists of an ordered arrangement of rows and columns. The tables are inaccessible as a scanned PDF component such as all other components in scanned PDF .Additionally there is no guarantee for tables to representing correct ordinary structured PDF due to lack of tags.

Table reader module extracts all table cells sort them based on column or row. The method to access each data cell individually is based on finding all columns and rows intersection points To find these points Table Reader uses several image processing techniques. Since table structure often contains vertical lines as column separator and horizontal lines as row separator using morphology erode and dilate technique first removes vertical lines and provides Rows position this processes repeated by removing horizontal lines and obtain columns position

Now by using crop technique and all intersection points table is segmented to cells .All cell are tagged based on positions and sent to OCR .Presenting cell segments to vision impaired users is the main issue regarding table accessibility. Listening to table straight through, without chance to see it visually can be quite confusing. Even by seeing table contents, it can still be confusing if the table is not marked up properly. It means table content linearization is not

enough for tables efficient usability To provide data table accessibility and usability, this access must use both passive and active methods. It means Table Reader must have navigation ability and user can interact with data.

(PK, FK)	Column Name	Description	Data Type	Size	Required	Default value
PK	CustomerID	Unique five-character code based on customer name.	Text	5	Yes	
	CompanyName	Customer Company	Text	40	Yes	
	ContactName	Name of the person who should be contacted	Text	30		
	ContactTitle	Title of the contact person	Text	30		
	Address	Street or post-office box	Text	50	Yes	
	City	City name	Text	20	Yes	
	State	State Code	Text	2	Yes	
	Zip	Five digit zip code	Text	5	Yes	
	Phone	10 digit phone number	Text	15	Yes	
	Fax	Fax number	Text	15		
	Email	Customer email address	Text	20		

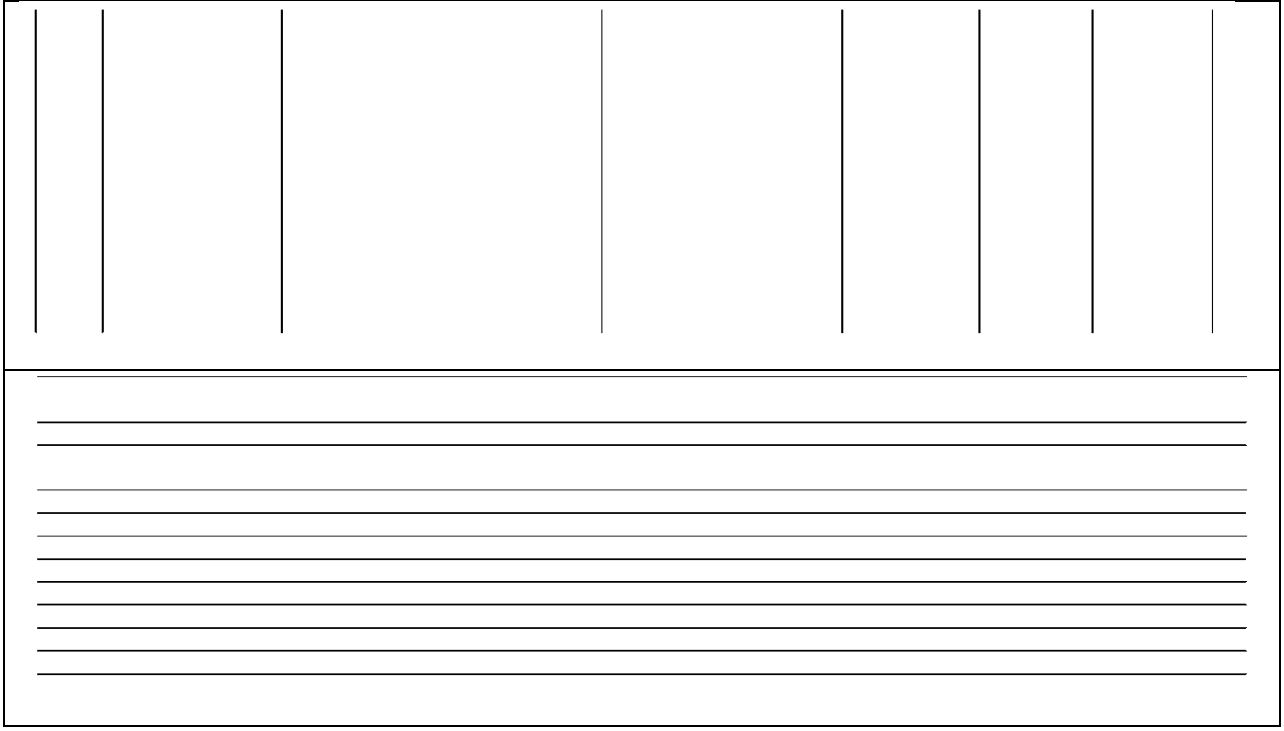


Figure 4. A sample table and its vertical and horizontal separator lines obtained by morphology technique

6 CONCLUSION

.The research described in this paper has undertaken to develop an application prototype to make HVTO accessible to vision impaired users. The flowchart shown in Figure 5 clearly labels the steps that are included in this process. In developing this prototype, several Open Source software packages have been used as a baseline. These packages include: espeak, ImageMagick, OCRopus. These have then been used with existing algorithms as well as further developed with the custom algorithms in an attempt to be able to produce accessible HVTO to visually impaired people.

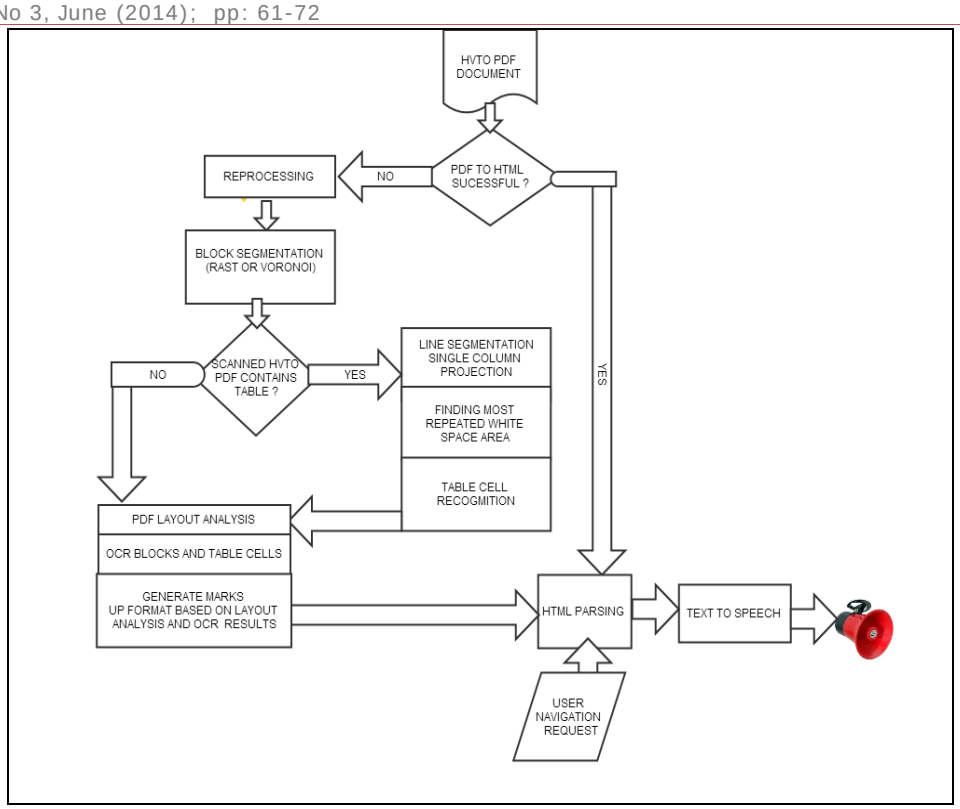


Figure 5. HVTO accessibility application flowchart

REFERENCES

[1]. Is PDF an appropriate choice for high volume transactional production printing? Crawford Technologies Inc. www.crawfordtech.com

[2]. P, Ganza, (2010), Accessibility and the High-volume Transaction Output Enterprisenn

[3]. T.M Breuel,"The OCRopus Open Source OCR System" doi=10.1.1.99.8505C

[4]. A Nazemi, I Murray, D McMeekin(2014) Layout Analysis for Scanned PDF and Transformation to the Structured PDF Suitable for Vocalization and Navigation Computer and Information Science 7 (1), 162

[5]. K, Kise. ,A, Sato,. and M, NB Iwata "Segmentation of Page Images Using the Area Voronoi Diagram." Computer Vision and Image Understanding, vol. 70(3), June 1998, pp. 370-382.

[6]. A,Winder. (2010),"Extending the page segmentation algorithms of the OCRopus documentation layout analysis system", Boise State University Graduate College, Theses and Dissertations. Paper 122.

[7]. T.M ,Breuel,. "Two Geometric Algorithms for Layout Analysis." Document Analysis Systems, August 2002, pp. 188-199

TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 2, ISSUE 3
ISSN: 2054 - 7390

Multidimensional Multi-granularities Data Mining for Discover Association Rule

Johannes K. Chiang[1], Chia-Chi Chu[2]
Department of Management Information Systems, Cloud Computing and Operation Innovation Center, National Chengchi University, Taipei, Taiwan
[1]jkchiang@nccu.edu.tw; [2]102356020@nccu.edu.tw

ABSTRACT

Data Mining is one of the most significant tools for discovering association patterns for many knowledge domains. Yet, there are deficits of current data-mining techniques, i.e.: 1) current methods are based on plane-mining using pre-defined schemata so that a re-scanning of the entire database is required whenever new attributes are added. 2) An association rule may be true on a certain granularity but false on a smaller ones and vise verse. 3) Existing methods can only find either frequent rules or infrequent rules, but not both at the same time.

This paper proposes a novel algorithm alone with a data structure that together solves the above weaknesses at the same time. Thus, the proposed approach can improve the efficiency and effectiveness of related data mining approach. By means of the data structure, we construct a forest of concept taxonomies which can be applied for representing the knowledge space. On top of the concept taxonomies, the data mining is developed as a compound process to find the large-itemsets, to generate, to update and to output the association patterns that can represent the composition of various taxonomies. This paper also derived a set of benchmarks to demonstrate the level of efficiency and effectiveness of the data mining algorithm. Last but not least, this paper presents the experimental results with respect to efficiency, scalability, information loss, etc. of the proposed approach to prove its advantages.

Keywords: Multidimensional Data Mining, Granular Computing, Concept Taxonomy, Association Rules, Infrequent Rule, information Lose Rate

1 INTRODUCTION

While Service Innovation is getting more interests in scientific and business communities, Data Mining turns out to be increasingly important for knowledge discovery of innovative services. Association rules can be used to figure out simple yet useful insights on services [5, 13, 17]. Significant examples are finding new purchasing behaviors for shops and new portfolios of rationale services. For example, “52% of the customer those buy product X also buy product Y”.

DOI: 10.14738/tmlai.23.259
Publication Date: 9th June 2014
URL: http://dx.doi.org/10.14738/tmlai.23.259

Given such association rules, we can decrease the costs of the product X, and raise the quality level of product Y to make more benefits.

However, most conventional data mining approaches only perform a plane scan over the databank based on a predefined schema for searching. Questions often arise such as: Should there be any other influencing factor like W on purchasing of product Y taken into account? Since most association rules apply in a context of certain breadth, the knowledge usually exists in multidimensional insides [5]. In the meantime, adding attributes to the data warehouse is meant to change the schema and initiates a full re-scan that would consume extra time.

The second problem of the conventional mining approaches lies in the assumption that the rules derived should be effective throughout a data warehouse as a whole. Nevertheless, this obviously is not true in real-life cases [5]. Different association rules can be found in different segments of the database. If a mining tool deals only with the database as a whole, the meaningful rules that are only partially true will be overlooked.

The goal of this research is to develop an approach with novel data structure and efficient algorithm for the multi-dimensional data mining for association patterns in various granularities. The crucial issue is to explore association patterns with more efficient and accurate multidimensional mining for the association patterns on different granularities. And, the data mining approach has to be very flexible and robust.

2 BASELINE OF THE RESEARCH

2.1 Multidimensional Data Mining

Finding association rules efficiently involving multi-attributes is an important subject for data mining. Association Rule Clustering System (ARCS) was proposed in [10], where association rule clustering is proposed for a 2-dimensional space. The restriction of ARCS is that it generates only one rule at a time of clustering. Subsequently, it takes massive redundant scans to find all rules.

The method proposed in [16] mines all large itemsets first and then applies a directed graph to assign attributes according to the priorities given by user for each attribute. Since the method is meant to discover large itemsets over a database as the whole, certain infrequent rules may be lost due to several granularities. Different priorities of the condition attributes will infer different rules so that user has to try with all possible priorities to discover all possible rules.

2.2 Frequent and Infrequent Rules

Records in a transactional database contain simple items identified by Transaction IDs using conventional methods. The notion of association is applied to capture the co-occurrence of items in transactions. There are two important factors for association rules: support and confidence. Support means how often the rule applies while confidence refers to how often the

rule is true. We are likely to find association rules with high confidence and support. Some data mining approaches allow users to set minimum support/confidence as the threshold for mining [6, 10]. Efficient algorithms for finding infrequent rules are also in development.

2.3 Apriori Algorithm

2.3.1 Apriori Algorithm

The Apriori algorithm is a level-wise iterative search algorithm for mining frequent itemsets w.r.t association rules [1, 3, 5, 7, 13, 14, 17]. The key drawback of the Apriori algorithm is that it requires k passes of database scans when the cardinality of the longest frequent itemsets is k. In addition, the algorithm is computation intensive in generating the candidate itemsets and computing the support values, especially for applications with very low support threshold and/or vast amount of items. In this algorithm, if the number of first itemsets element is k, the database will be scanned k times at least. So, it is not efficient enough. The key point for improving the algorithm is to reduce the number of itemsets.

2.3.2 AprioriTID Algorithm [9]

The AprioriTID is a variant of the aforementioned Apriori algorithm which reduces the time needed for the frequency counting procedure by replacing every transaction in the database by the set of candidate sets that occur in that transaction [9]. This is done by iterating each candidate sets repeatedly.While the AprioriTID algorithm is much faster in later iterations, it is much slower than original Apriori in early iterations. This is mainly due to the additional overhead that is created when the adapted transaction database Ck does not fit into main memory and has to be written into disk [4]. If a transaction does not contain any candidate k-sets, then Ck will not have an entry for the transaction. Hence, the number of entries in Ck may be smaller than the number of transactions in the database, especially at later iterations of the algorithm. Other drawbacks of ApririTID are that the database modified by Apriori-Gen can be much larger than the initial database and only faster in the later stages of the scans.

2.4 Concept Description and Knowledge Taxonomy

The issues of data structures and concept description models for data mining when comparing works dealing with algorithms are less discussed till. The concept description task is problematic, since the term “concept description” is used in quite different ways in related discussions. In this situation, researchers argue for a de facto standard definition for the concept description [8, 18]. At this beginning stage, it is easier to deal with normal criterion on higher abstraction level for the concept description, such as comprehension [8] and compatibility [4].

Researchers view concept description as a form of data generalization and define the concept description as a task that generates descriptions for the characterization and comparison of the data [8]. Similar concept appears in the development of ontology for Semantic Web/GRID. Semantic Web can be described as an extension of the existing Web

where information is considered with priori well-defined meaning, enabling computer and people to work in cooperation centric to Internet [11]. The objective of such techniques is to enhance ill-structured content so that it can be interpreted universally by machines or humans.

In practical applications, ontology provides a vocabulary for specific domains and defines the meaning of the terms and relationships between them. In this paper, ontology refers to the shared understanding (comprehension) of domains of interests which is often conceived as a set of concepts, relations, axioms etc. Hence, the term "Taxonomy" is hereby similar to "Ontology" and both terms can be used to denote the classification or categorization of concepts that describe entities and relations among them. This paper applies the term Taxonomy rather than Ontology because the former is more flexible and even can cover the case with no semantic meaning.

3 METHODOLOGY

3.1 Representation schema and data structure

As mentioned in section 2.4, the issues of data structures regarding descriptive models are less discussed when comparing R&D works dealing with data mining algorithms. Therefore, we will present the building blocks of our representation schema and data structure, which are namely (1) Taxonomy, (2) Forest of Concept Taxonomies and (3) Association Rules.

For the sakes of comprehension and compatibility, we use the forest structure consisting of Concept-Taxonomies to represent the overall searching space, i.e. the set of all the propositions of the concepts. On top of this structure, the sets of association patterns can be formed by selecting concepts from individual taxonomies. The notions can be clarified with examples as follows:

3.1.1 Taxonomy

A category consists of domain concepts in a latticed hierarchical structure, while each member per se can be in turn taxonomy. An Example (see Figure 1) for customer's characteristics can be [Age, Sex, Marry, Urbanization], while for instance the taxonomy of Sex can [Male, Female] and Marry can [Married, Unmarried] so on.

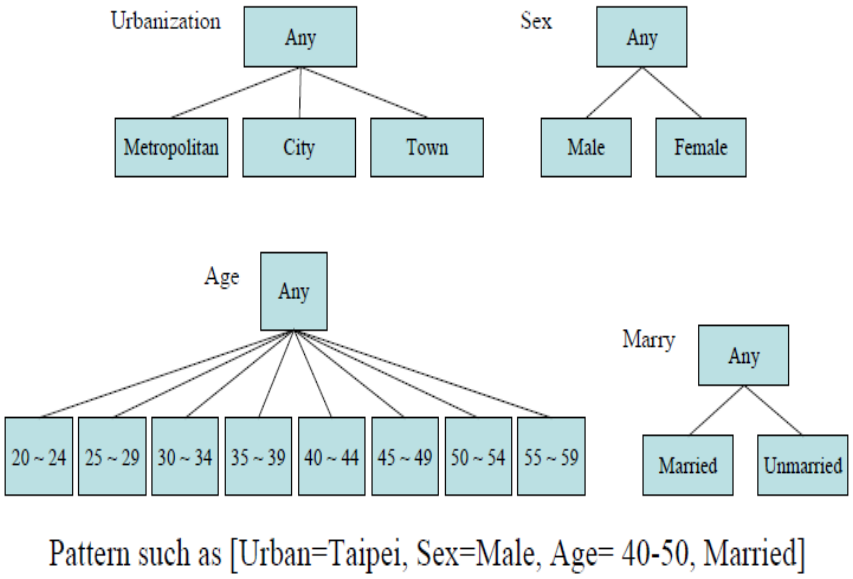


Figure 1: An Example for Forest of Concept Taxonomies

3.1.2 Forest of concept taxonomies

A hyper-graph for representing the universe of discourse or the closed-world of interests is built with taxonomies under consideration. An example of forest of taxonomies with respect to the location and Sex of customers is shown in Figure 2 below:

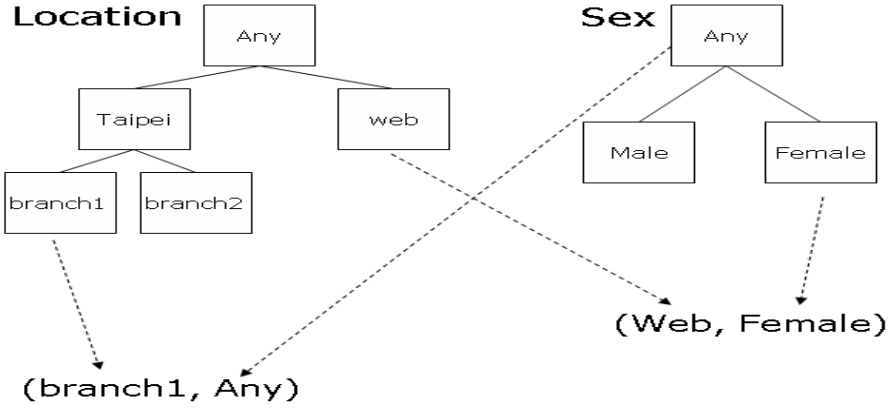


Figure 2: Examples of Forest Concept Taxonomies

3.1.3 Association Rule

An association rule typically refers to a portfolio's pattern which consists of elements taken from various concept taxonomies such as [(Location=branch1), (Sex=female)]. It owns support and confidence greater than the user-specified minSup and minConf respectively [4].

3.1.4 Element patterns and generalized patterns

An element pattern is composed of dimension atoms. On the other hand, if at least one of them is a dimension compound which combine several dimension atoms, we call this pattern a

generalized pattern. For example, <web, Female> is an element pattern, <branch1, Any> is a generalized pattern, and both them are multi-dimension patterns. We use to denote the i-th element pattern, and use to denote the j-th generalized pattern.

By the proposed multidimensional data mining of association rules, the notion of relation will be implemented by the belonging relationship between elementary patterns and generalized patterns rather than the semantics [4]. Other notations to be used in the following text are shown in Table 3 below:

Table 1: Concepts and Notations

Notation	Meaning
CT	Concept Taxonomy
Ei.	The *i*-th element segment
T[*Ei*]	an element segment over Ei in MD
Gi	The *j*-th generalized pattern
T[*Gj*]	The *j*-th combined segment over G
R*E* i	Rules w.r.t the *i*-th element segment
R*Gj*	Rules w.r.t the *j*-th generalized pattern
(*Gj* , *r*)	association rules over *Gj* w.r.t to match ratio *r*

3.2 The Multidimensional Multi-granularity data mining algorithm

The proposed data mining process can be formulated essentially with two cascading steps: (1) finding all itemsets in each elementary segment and (2) updating all combinations of the segments by the output of Phase 0. For the practical reason, the algorithm in Phase 0 can be replaced by any tool available elsewhere such as the Apriori algorithm so that an easy realization the phase 0 and then a segregation of the two steps enable the flexible mining on a distributed environment like Cloud and Grid. Figure 3 illustrates outline of the proposed algorithm extending the mining process into four phases.

```
1)  Input:
2)      Multidimensional Transaction Database MD
3)      Concept taxonomies for each dimension: CTx(X= 1-n)
4)      User given threshold: minsup, minconf, match ratio m
5)  Procedure:
6)      Phase0:
7)          to generate all Ei and Gj by CTx (x = 1 to n);
8)          build the pattern table;
9)      Phase1:
10)         For all Ei ⊂ G
11)             to discover all association rules r in T[Ei] as R_Ei;
12)     Phase2:
13)         for all Ej
14)             for all Gj that Ei ⊂ Gj
15)                 to update R_Gj using R_Ei;
16)     Phase3:
17)         for all Gj
18)             For all r (which satisfy m) in R_Gj
19)                 output (G_j, r);
20)     Output:
21)      all multidimensional association rules(p, r)
```

Figure 3: Outline of the proposed algorithm.

Outline of the proposed algorithm is shown in Figure 3. The input of the mining process involves 5 entities, namely (1) a multidimensional transaction database MD which is optional when a default MD is assigned, (2) a set of concept taxonomies for each dimension (CTs), (3) a minimal support, viz. minSup, (4) a minimal confidence, viz. minConf, and (5) a match ratio m for the relaxed match. The output of the algorithm encompasses all multi-dimensional associations with respect to the fully-relaxed match within the MD. The last three settings can help with finding frequent or infrequent rules.

The most significant feature of the algorithm is it's capability to discover both frequent and infrequent associations rules R_{Ei} (based on different levels of granularities) in the element segment T[E*i*] for each element pattern E*i*. After it, R_{Ei} is used to update R_{Gj}, i.e. the set of association patterns for every generalized pattern G*j* which includes E*i*. The heuristic regarding each element pattern is to find the large-itemsets per se and acknowledge its super generalized patterns with the result. The task of each generalized pattern is to decide which rules hold within it, according to the acknowledgements from the element patterns. The mining procedure needs only to work on each element segment to determine which rules hold in the compound segments. Thus, it is not necessary to scan all of the potential segments for finding the rules.

3.3 Pattern Generation and the Pattern Table

Being a pre-processing mechanism, the algorithm generates at first all elementary and generalized patterns with the given forest, where a pattern table for recording the belonging relationship between the elementary and generalized patterns is built. Given a set of concept taxonomies, a multi-dimensional pattern can be generated by choosing a node from each of the taxonomy. The compound of different choices represents all the multidimensional patterns.

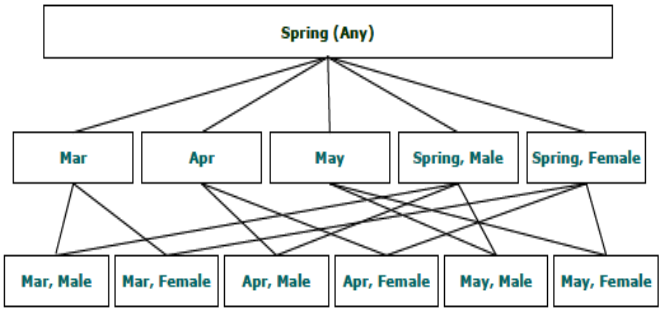


Figure 5: Belonging relationships between patterns

	(Mar)	(Apr)	(May)	(Spring, Male)	(Spring, Female)	(Spring)
(Mar, Male)	1	0	0	1	0	1
(Mar, Female)	1	0	0	0	1	1
(Apr, Male)	0	1	0	1	0	1
(Apr, Female)	0	1	0	0	1	1
(May, Male)	0	0	1	1	0	1
(May, Female)	0	0	1	0	1	1

Figure 6: The pattern table (for the relations in Figure 4) shows an example of the belonging relationship between 12 patterns in a lattice structure. The relationships are recorded in the form of bit map as shown in Figure 5 which includes element patterns and generalized patterns. In the table, a "1" indicates that the element pattern belongs to the corresponding generalized pattern and "0" indicates the case vice versa.

3.4 Update process

1) **for** all R_{Ei}
2) **for** all $G_j \supset E_i$
3) **if** (R_{Gj} never be updated)
4) $R_{Gj} = R_{Ei}$;
5) **else**
6) $R_{Gj} = R_{Gj} \cap R_{Ei}$;

Figure 6: The "Update" algorithm for the full match

In order to be more optimization algorithm, we proposed full match and relaxed match method for update process. After all patterns and the pattern table have been generated, the procedure reads the transactions of each element segment and then discovers all the association rules. The output of this phase is all R_{Ei} for each element pattern E*i* that will be fed as the input to the next phase for updating each R_{Gj} using R_{Ei}. For a full match illustrated in Figure 6, the update is done by intersection of the set R_{Gj} and the set R_{Ei}, where E*i* belongs to G*j* , let $R_{Gj} = R_{Ei}$ if R_{Gj} is updated for the first time. After all the intersections, the association pattern r left in R_{Gj} holds in all element segments covered by T[G*j*].

```
1)  for all R_Ei
2)      for all G_j ⊃ E_i
3)          for all r in R_Ei
4)              if (r ∉ R_Gj)
5)                  add r to R_Gj;
6)                  R_Gj.r.count = 1;
7)              else
8)                  R_Gj.r.count++;
```

Figure 7: The "Update" procedure for the relaxed match

For the relaxed match as shown in Fig. 7, a counter for each rule in R_{Gj} is set. While using R_{Ei} for updating R_{Gj}, the counters of both R_{Gj} and R_{Ei} are incremented by one and the rules, those appear in R_{Ei} but not in R_{Gj}, will be added to R_{Gj} while setting the counter to one. After all the update process, the association rule r in R_{Gj} whose counts exceed m|T[G*j*]| holds in at least m *100% of the element segments T[E_i] that are covered by T[G*j*], and thus (G*j*, r) is a multidimensional association rule for the relaxed match in MD.

Full match can ensure that all association rule be found in various granularities. But, it may be too restrictive to ignore some rules. On the other hand, relaxed match can solve "restrictive" problem and hold more association rules which may be our interesting rules. User can adjust the m ratio which ranges between 0 and 1.

For example, suppose we have a generalized segment <Spring> which covers three element segment <March>, <April>, and <May>. Finding patterns of each element segment <March>{A},{B},{C}、<April>{B},{C} and <May>{B},{E}. As we above-mentioned algorithm that update each R_{Gj} using the R_{Ei} come from previous phase. For the full match case, we just can hold rule B in <Spring> generalized segment R_{Gj} because only rule B exists every element segment R_{Ei}. For the relaxed match case, we suppose m = 0.6 (result of count numbers should greater than 1.5 times) and count numbers of all rules in each element segment R_{Ei}: {A=1}、{B=3}、{C=2}、{E=1}. Hence, we hold rule B and C in <Spring> generalized segment R_{Gj}.

3.5 The Output Function

For a full match, the algorithm outputs all the (G*j*, r) pairs for every r left in each R_{Gj}. For a relaxed match, it outputs all the (G*j*, r) pair for every r in each R_{Gj} where the count exceed |mT[G*j*]|. By means of this approach, loss of finding the rules that only hold in some segments can be prevented. And, pickup of multidimensional association rules that do not hold over all the range of the domain can also be avoided. For example, the full match can guarantee that the corresponding rules, those hold only in two months of spring but fail in the rest one, will never be counted as an association rules with respect to whole spring.

3.6 The Breakthroughs for Incremental Data Mining

A breakthrough hereby is that the incremental data mining can be realized with the proposed approach. By keeping out the rules deduced in each element segment, we only need to search the new data. That is, using the proposed approach, we can produce the new association rules by combining the rules discovered from the new data with existing rules to reduce redundant scan on the old data. The following section will present our experimentation results.

3.7 Design of metrics for measuring data mining

In order to assure the performance, we need to design metrics for measuring the mining performance, at least to measure whether it is better than the prior algorithms. By cascade evaluating the results of a hypothetical measurement, we can evaluate the consequence from any sequence of measurements to determine the optimal next measure. For this reason, a one-step look-ahead strategy based on Shannon's Entropy Function is adopted and the capacity of ICT systems can be described in the following form [4, 15]:

$$C = B * [\log2\ (1 + S/N)] \tag{3}$$

where B is the bandwidth, (S/N) is Signal-to-Noise(S/N) ratio.

Drawing on this equation, the function for the performance of data mining can be formulated as follow:

C=|D| [log2 (1+information lost ratio)], where |D| is the number of transactions in whole transaction database [4].

While WSE*i* denotes each element segment in the measure, the WSE*i* of an element segment T[E*i*] can be generated by a uniform distribution between 0 and SM. Suppose there are N element segments, the number of transactions in the element segment T[E*i*] is:

$$|D_{Ei}| = \frac{|D|}{\sum_{a\to 0}^{n} WS_{Ea}} WS_{Ei} \tag{4}$$

Thereafter, the definitions of information loss are:

$$\text{discrete ratio} = \frac{|\{r \mid r \text{ holds in } T[Gj] <Gj,r> \text{ doesn't hold in } \mathbf{MD}\}|}{|\{r \mid r \text{ holds in } T[Gj]\}|} \tag{5}$$

Definition 1: discrete ratio is the ratio of the number of rules pruned by the improved algorithm to the number of rules discovered by prior mining approaches.

$$\text{lost ratio} = \frac{|\{<\text{Gj,r}> \mid <\text{Gj,r}> \text{ holds in MD r doesn't hold in T[Gj]}\}|}{|\{<\text{Gj,r}> \mid <\text{Gj,r}> \text{ holds in MD}\}|} \tag{6}$$

Definition 2: lost ratio is the ratio of the number of rules discovered by the improved algorithm but lost in the previous mining approaches to the number of rules discovered by the improved algorithm.

4 EXPERIMENT AND EVALUATION

4.1 Experiment scenario

To measure and prove the performance of the method, a scenario for a wholesale business using synthetic data are established for the test. The wholesales enterprise runs various business branches and a web-site for its operations. Data from four branches and the website are gathered for the experiment. We take five of the various attributes (Abode, Sex, Occupation, Age and Marriage) as the dimensions for the test. Adding with the product catalog and price/profit record, there are 7 dimensions and we build the concept taxonomies for each dimension.

Table 2: Three Types of Experimental Data Set

Type1	To generate a single set of maximal potentially large itemsets and then generate transactions for each element pattern Ei following apriori-gen.[3]
Type2	Diagnosis-2, Therapy1. Beside a set of common maximal potential large itemsets, to generate maximal potentially large itemsets for each element pattern Ei. and then generate transactions for each element pattern Ei and the common maximal potentially large itemsets respectively following the apriori-gen[3]
Type3	generating a set of maximal potentially large item-sets for each element pattern Ei, and then generating transactions for each element pattern Ei from its own maximal potentially large itemsets following the apriori-gen.[3]

The test bench is implemented with Java on a PC Server with an AMD processor and the data mining software is implemented with Java. Data from different branches and the website are collected for our experiment. To examine the effect of different customer behaviors, we generate three data types as illustrated in Table 2. The parameters and the default values of the data sets are illustrated in Table 3. There are 118 multidimensional patterns from these taxonomies, 44 of them are element patterns and the other 74 of them are generalized patterns. The mining tool should find all large item sets for the 74 generalized patterns.

Table 3: Parameters and default values of data sets

Notation	Meaning	Default
$\|D\|$	Number of transactions	100K
$\|T\|$	Average size of transactions	6
$\|I\|$	Average size of maximal potentially large itemsets	4
$\|L\|$	NBumber of maximal potentially large itemsets	1000
N	Number of items	1000
S_M	The maximum size of segmentation	50

4.2 The Results of Experiment

At first, the 74 generalized patterns are successfully found. The key feature of the algorithm as illustrated in Figure 9 is that it is linear (and hence highly scalable) to the number of records and that it is flexible in terms of reading various data types. The test result w.r.t scalability in

Figure 9 illustrates that the algorithm takes execution time linear to the number of transactions of all three data types. The experiment results of both the test (see Figure 8 and 9) illustrates that the new algorithm is superior to conventional methods in several areas:

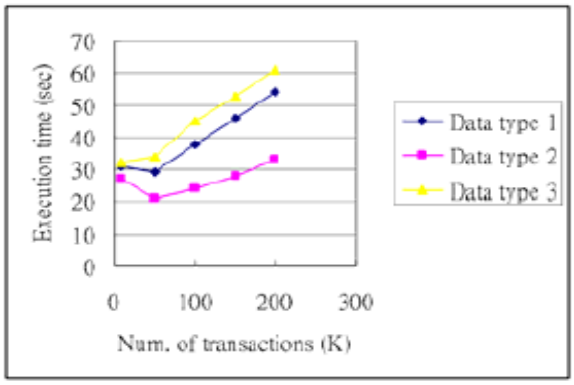


Figure 8: Scalability test w.r.t. the no. of transactions

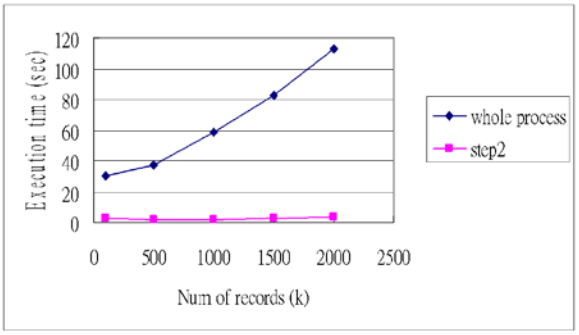


Figure 9: Scalability experiment w.r.t. the no. of records

Execution time with regards to number of transactions is linear for the data types tested for the whole process. This means that the time and space cost of executing our algorithm do not increase exponentially as compared to conventional methods.

Phase 2 (the update phase) of our algorithm is an important space and time saver as illustrated by the Figure 8; execution time is also linear and time taken to read up to 2000k records took less than 5 seconds. This means that data patterns from new data can be quickly extracted and used to update the existing pattern table for immediate use.

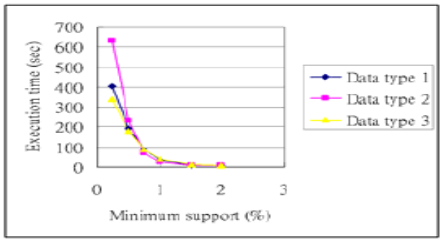


Figure 10: Efficiency in Relation to Minimum Support

In general, an increase of element patterns with result in an increase in execution time; the key to scalability is having the execution time increasing in a linear manner with an increase in element patterns. In Figure 11, all three data types experienced an increase of execution time with an increase of element pattern in a linear fashion, thus making our algorithm efficient.

Most importantly, an increase in element patterns leads to a less than proportion increase in execution time, making out the algorithm highly scalable. Reading off Figure 10, a 4 time increase of 30 element patterns from 10 to 40 will result in:

- 75 times increase in execution time for data type 1 from 20 seconds to 35 seconds.
- 1.67 times increase in execution time for data type 2 from 15 seconds to 25 seconds.
- 2.05 times increase in execution time for data type 3 from approximately 22 seconds to 45 seconds.

The impact of minSup on the algorithm can be categorized in terms of efficiency, discrete ratio and lost ratio. All of such algorithms are sensitive to the minimum support; the smaller the minimum support, the longer the execution time. However, we have shown that the real execution time of the step 2 (the update) in the proposed algorithm is relatively much shorter than the whole process (see Figure 8).

The test results proved that an increase in minSup will lead to greater returns of investment in terms of time efficiency; this is in line with one of the core objectives of building an efficient algorithm. Our algorithm is more efficient than conventional methods in terms of execution time over data. For instance in Figure 11, a 10 time increase (from 0.1 to 1) in minSup leads to a more than proportionate decrease in execution time across all data types:

- Execution time for data type 1 decreased by approximately 10 times, from approximately 400 seconds to approximately 40 seconds in terms of execution time.
- Execution time for data type 2 decreased by more than 30 times, from more than 600 seconds to approximately 20 seconds in terms of execution time.
- Execution time for data type 3 decreased by more than 11 times, from approximately 350 seconds to approximately 30 seconds in terms of execution time.

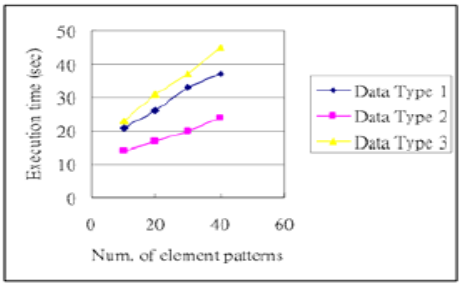


Figure 11: Efficiency in Relation to Minimum Support

The discrete ratio is the ratio of the number of rules pruned by the proposed algorithm to the number of rules discovered by prior mining approaches. Figure 12 illustrates the ratio of rules pruned by the proposed algorithm against minSup. In general, all three data types (except

for data type 1) exhibited an increase of ratio with an increase of minSup from approximately 0.2% to 2%.

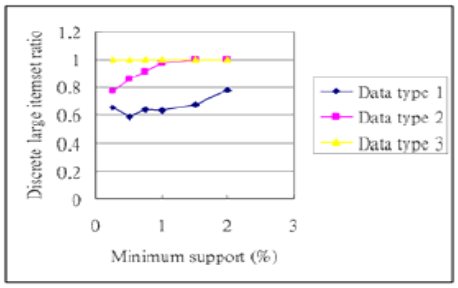


Figure 12: Effects of MinSup on discrete large itemsets ratio

The test results point the fact that the proposed algorithm can effectively decrease unwanted generalized patterns in which elemental data patterns is not true. This greatly helps users to focus on data patterns that are useful for their organizations while uncovering niche data patterns. For instance with a higher setting value, only <Female, Age 30-50, buy SK-II > will be found instead of <Age 30-50, buy SK-II>.

Figure 13 illustrates the test result on lost ratio, i.e. the influence of minSup values on the lost rules by other mining tools in comparison to this approach. All three data types experienced an increase in lost ratio over an increase in minSup from 0.25% to 2%, with the greatest increase in data type 2, followed by data type 3 and finally data type 1.

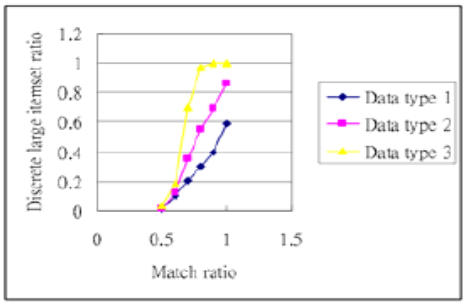


Figure 13: Effects of match ratio on discrete large itemsets ratio

The test results prove that the proposed algorithm will help users uncover useful data patterns which otherwise would be uncovered by traditional approaches. Thus, our objective of uncovering niche data patterns that would otherwise be left out is met and proved by this test result.

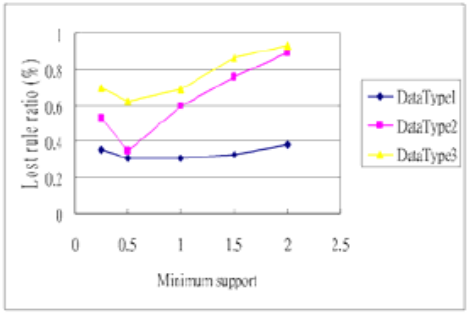


Figure 14: Effects of match ratio on lost itemsets.

Increasing the match ratio would decrease unwanted data patterns in general. Figure 13 illustrates the effect of match ratio (r) on discrete ratio.Similar to the above test results, an increase of m from 0.5 to 1 results in a more than proportional increase in discrete ratio across all three forms of data types. The significance of this test result is congruent with the test results above; the algorithm is efficient and scalable without losing flexibility and helps uncover niche data patterns.

5 SUMMARY

The paper proposes an approach including a novel data structure and an efficient algorithm for mining association rules on various granularities. The advantages of this approach over existing approaches include (1) more comprehensive and easy-to-use (2) more efficient with limited scans (3) more effective with finding rules hold in different granularity levels (4) capable of finding frequent patterns and infrequent patterns while users can choose the full match and the relaxed match (5) low information loss rate (6) capable of incremental mining of association rules to avoid unnecessary re-scan.

The whole development process and experimental measurement of the multidimensional data mining approach were discussed in this paper. The test result reveals that its performance, efficiency, scalability and information loss rate are better than the current approaches. The effects of perceived issues and potential development of data mining and big data strategy are worthy of further investigation. And, deployment of big data mining over the MapReduce on top of cloud computing architecture is our target of our future research.

REFERENCES

[1]. R. Agrawal and J. C. Shafer (1996). "*Parallel Mining of Association Rules,*" IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, pp. 962-969.

[2]. R. Agrawal and R. Srikant (1994). "*Fast Algorithms for Mining Association Rules in Large Databases*," in Proceedings of the 20th International Conference on Very Large Data Bases.

[3]. R. Agrawal, T. Imielinski and A. N. Swami (1993). "*Mining Association Rules between Sets of Items in Large Databases,*" in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data.

[4]. J. K. Chiang (2007). "*Developing an Approach for Multidimensional Data Mining on various Granularities ~ on Example of Financial Portfolio Discovery,*" in ISIS 2007 Proceedings of the 8th Symposium on Advanced Intelligent Systems, Sokcho City, Korea.

[5]. J. K. Chiang and J. C. Wu (2005). "*Mining Multi-Dimension Rules in Multiple Database Segmentation-on Examples of Cross Selling,"* in Proceedings of the 16th International Conference on Information Management, Taipei, Taiwan.

[6]. T. M. Cover and J. A. Thomas (2006). *Elements of Information Theory,* 2nd ed., Wiley.

[7]. R. Feldman and J. Sanger (2007). *The Text Mining Handbook – Advanced Approaches in Analyzing Unstructured Data,* Cambridge University Press.

[8]. J. Han and M. Kamber (2006). *Data Mining - Concepts and Techniques,* 2nd ed., Morgan Kaufman.

[9]. L. J. He, L. C. Chen and S. Y. Liu (2003) "*Improvement of AprioriTid Algorithm for Mining Association Rules*," Journal of Yantai University(Natural Science and Engineering Edition), vol. 16, no. 4.

[10]. B. Lent, A. Swami and J. Widom (1997). "*Clustering Association Rules,*" in Proceedings of the 13th International Conference on Data Engineering.

[11]. M. Li and M. Baker (2005). The GRID – Core Technologies, Wiley.

[12]. B. Liu, W. Hsu and Y. Ma (1999), "*Mining Association Rules with Multiple Minimum Supports,"* in Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[13]. G. Shmueli, N. R. Patel and P. C. Bruce (2007)."*Association Rules*," in Data Mining for Business Intelligence, Concepts, Techniques, and Applications, Wiley, pp. 203-215.

[14]. R. Srikant and R. Agrawal (1995). "*Mining Generalized Association Rules,*" in Proceedings of the 21th International Conference on Very Large Data Bases, Zurich, Switzerland.

[15]. W. Stallings (2004). "*Channel Capacity*," in Business Data Communications, 6th ed., Pretice Hall, pp. 470-471.

[16]. P. S. Tsai and C. M. Chen (2004). "*Mining interesting association rules from customer databases and transaction databases,*" Information Systems, vol. 29, no. 8, p. 685–696.

[17]. C. Vercellis (2009). "*Association Rules*," in Business Intelligence, Data Mining and optimization for Decision Making, Wiley, pp. 277-290.

[18]. The CRISP-DM Consortium, CRISP-DM 1.0 (2000), www.crisp-dm.org.

The Author

Prof. Dr.-Ing. Johannes K. Chiang is now a faculty member of the Department of MIS and the Deputy Director of the Center for Cloud Computing and Operation Innovation at National Chengchi University Taipei. He received his academic degree of Doctor in Engineering Science (*Dr.-Ing., Summa Cum laude*) from the RWTH University of Aachen Germany. His current research interests include Cloud Computing, Semantic Web, Business Intelligence, Data Mining, e-Business and ebXML. He also serves as a consultant for several government agencies in Taiwan and as an active member of various international affiliations, such as IEEE, ACM, CSIM and ITMA etc. before 1995, he has been a research fellow at RWTH of Aachen and a Manager of EU/CEC ESPRIT Programs

TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 2, ISSUE 3
ISSN: 2054 - 7390

An Android Malware Detection Architecture based on Ensemble Learning

Mehmet Ozdemir, Ibrahim Sogukpinar
Department of Computer Engineering, Gebze Institute of Technology, Gebze, Kocaeli, Turkey;
mehmet.ozdemir@tubitak.gov.tr, ispinar@bilmuh.gyte.edu.tr

ABSTRACT

In the scope of anomaly based Android malware detection, different type of features has been used to represent applications and lots of algorithms have been applied to evaluate these features. Although researchers have reported accurate results, in order to improve accuracy, sensitivity and generalization, we suggest using an ensemble learning approach for Android malware detection. In this study, we propose to use an ensemble learning system whose base learners are built with different feature subsets which are extracted and processed with multiple methods, and selected with a proposed selective ensemble approach which is based on three criteria: Accuracy, sensitivity and diversity.

Keywords: Ensemble Learning, Multiple Classifier Systems, Mixture of Experts, Selective Ensemble, Malware Detection, Android Malware.

1 INTRODUCTIN

Detection of malware using data mining techniques requires representing applications as features. These features are then used to build mining models which provide predictions about maliciousness of applications. However, selecting most suitable features among lots of feature types and processing the selected features with correct algorithms is not a simple task.

In order to represent applications completely, selected feature types must be carefully decided. Choosing few features types may not represent all applications sufficiently. For instance, network activity of an application can be recorded by executing that application and can be used as features in order to represent that application. However, it can't be guaranteed that these features will enough to distinguish malware and benign applications because there is a good chance of being malware without network activity. On the other hand, feature types which are complement of each other must be considered. For example, if API calls are used as

DOI: 10.14738/tmlai.23.261
Publication Date: 9th June 2014
URL: http://dx.doi.org/10.14738/tmlai.23.261

features, these calls must be gathered from both native code and byte code because it isn't known where the malicious intent would be in advance.

Analysis type of extracted features may be another complementing factor of features because static and dynamic analyses are generally considered as complement of each other. For instance, it is hard to gather useful information from obfuscated code via static analysis but dynamic analysis can reveal that code's behavior. On the other hand, dynamic analysis may not reveal some malware intentions (e.g. time bomb attacks) because of limited execution time but static analysis may catch some pattern about these kinds of intentions.

Choosing the mining algorithms are also important as well as chosen feature types. Once the features are extracted, feature selection algorithms are generally applied before providing these features to the learning algorithms. There are kinds of feature selection and learning algorithms whose assumptions are different from each other. None of these algorithms have been proven to be best for a specific problem and thus, several methods should be compared before making predictions. Additionally, comparison of chosen algorithms may be crucial. For example, two different learning models may have the same accuracy but their predictions may be totally different.

Contribution of this paper is twofold. First one is to show benefits of ensemble learning for malware detection problem. Second one is to increase accuracy and sensitivity of malware detection operation with proposed architecture. In order to do this, we propose to use an ensemble learning system whose base learners are generated using different kind of features which are extracted from different aspects of applications, and processed using different kind of selection and learning algorithms. After that, some of the generated base learners are selected with a proposed heuristic algorithm to improve accuracy and sensitivity. Finally, selected base learners are combined via stacking or majority voting to build an ensemble system.

The rest of the paper is organized as follows. Related works are described in Section 2. Overview of ensemble learning and feature selection is presented in Section 3. Architecture of proposed method is explained in Section 4. Experimental results and evaluation are given in Section 5. Conclusions and future work are described in the last section.

2 RELATED WORKS

Misuse and anomaly detection are two general approaches to reveal malware applications. In misuse detection, signatures are generated for each specific kind of malware and once a signature is generated, that specific malware can be found precisely [10]. However, this method fails to detect novel malware. On the other hand, anomaly based systems build a general model using an application dataset and applications that don't fit to this model are considered as anomalous [23]. Although novel malware can be detected by anomaly based systems, false predictions may be problem for this method.

Features are extracted from applications by applying two general approaches: Dynamic and static analysis. In dynamic analysis, applications are run on a device or emulator and behavior of the application or system is watched. System calls [4], sensitive data tracking [9], system logs [22], messaging & call information, CPU load etc. [23] are some dynamically extracted features. In static analysis, features are extracted without running the application. API calls [32], opcodes and operands [30], class hierarchies, used packages [32] and control flow graphs of applications [20] are some statically extracted features from compiled source of applications. Also it is possible to extract static features from Android application package (apk) content such as resource locations and Android permissions [32]. Also, meta information such as description, download count and price of application [26] can be statically extracted from marketplace.

Zhou and Jiang [31] collected 1260 malware samples and used these samples to test existing mobile security tools. Best accuracy rate was reported as 79.6%. After this study, researchers have reported better accuracies from their studies. However, sensitivities of some studies were considerably low. Sensitivity[1] is the measure of ability to identify positive (malware) samples correctly. It is very important for malware detection tools because labeling a malware application as benign (false negative) will cause malware application to be added to the application market, which is not acceptable, while false positives can be fixed by security experts of the application market. Hence, this situation motivated us to create a detection tool both accurate and sensitive.

3 OVERVIEW

3.1 Ensemble Learning

Ensemble learning[2] is a machine learning method which is used to improve accuracy of learning systems by *generating* a set of base learners and then *combining* outputs of these base learners. Effectiveness of ensemble learning has been proven in several studies [7] when the base learners have error rates less than random guessing and their errors are uncorrelated (i.e. base learners are diverse).

There are three stages for constructing ensemble systems [25]. First one is to generate accurate and diverse base learners. Diversity is a term which is used to define variation among outputs of learning models. It is an important factor for effective ensemble systems because if base learners are identical, obviously, combining them wouldn't provide any information gain.

Although the importance of diversity over effective ensemble systems is clear, there is not a common explanation for diversity in literature. Though, there are several investigations to explain diversity among learning models for regression and classification problems [3].

[1] Recall rate, true positive rate and hit rate are other terms used for sensitivity.

[2] This term is also known as multiple classifier systems, mixture of experts, committees of learners.

Diversity measures are one of the proposed methods to measure diversities among base learners and it fall into two categories, the pair wise measures and the non-pair wise measures. In the pair wise measures, firstly, diversity is measured for each pair of base learners and then diversity of the ensemble system is calculated by averaging these pair wise diversities. In the non-pair wise measures, diversity of the ensemble is measured directly without considering divergence between pair of base learners. In order to construct better ensemble systems, Kuncheva and Whitaker [16] compared several diversity measures for classification problems in terms of correct/incorrect outputs. As a result, they reported that existing divergence generation methods are valid but measuring diversity in order to build better ensemble systems is an open problem.

Since there is not a general acceptance about diversity, diverse base learners are tried to be generated intuitively with different methods including, using different learning algorithms [15], using different parameters of the same learning algorithm [13], using different subsamples in the training set [11], using different feature subsets of the training set [14], manipulating the output targets [6], injecting randomness into algorithms [17].

Selecting appropriate base learners is the second stage to construct ensemble systems. Instead of using all generated base learners, optionally, a subset of them is selected. Investigations show that such a stage improves the performance of the ensemble system [19].

There are two strategies for the concept of base learner selection as the "direct strategy[3]" and the "overproduce and choose strategy[4]". Direct selection of base learners is done internally by learning algorithm itself at the training phase such as selecting base learners with pruning functions in boosting. In overproduce and choose strategy, firstly, lots of base classifiers are generated and then a subset of them is selected by applying search algorithms or rules. These rules or search algorithms evaluate base learners based on one or more evaluation criteria like accuracy or diversity measure. Selecting best ones [18], selecting via heuristics [24], selecting with genetic algorithms [21] are some selection techniques in literature.

Final stage of constructing ensemble systems is to combine selected base learners. In order to combine base learners' outputs for final decision, a simple majority voting may be used (or a simple averaging for regression). Or, another learning model[5] may be constructed using base learners' outputs (i.e. stacking [28]). Actually, many combination schemes have been proposed which are based on different theories. Studies that convert class labels to continuous values and apply regression to the converted values for combining classifier outputs even exist. More

[3] This strategy is also known as dynamic selection, or, test and select methodology.

[4] Static selection or selective ensemble terms are also used for this strategy.

[5] This learning model is called as meta or strong learner.

details can be found Tulyakov et al.'s study which gives a comprehensive review of combination schemes for classification problems under different categorizations [27].

In conclusion, main purpose of ensemble learning is to improve generalization of learning models in general, not to find best accuracy for a specific dataset. Hence, accuracy of an ensemble system may be lower than its best base learner but commonly are expected to be higher than average. Though, ensemble systems' accuracies may be as high as their best base learner's. For instance, Saso and Bernard constructed an ensemble system using stacking scheme and they compared performance of their ensemble system with a best classifier chosen from a cross validation. They reported that results were comparable [8].

3.2 Feature Selection

Representing applications as features may produce very high dimensional data. Number of the features may be as high as tens of thousands. Running learning algorithms with this data would increase time and space complexity. Instead of taking into account all of the produced features; a subset of them is selected using a feature selection algorithm. Several studies showed that, using such an algorithm may remove irrelevant and redundant feature and also increase accuracy [2].

Guyon and Elisseeff presented a detailed review about feature selection [12]. Although feature selection for unsupervised learning was mentioned, main focus of their study is supervised learning. In the following part, feature selection for supervised learning will be discussed briefly.

There are two general approaches for feature selection algorithms: The *wrapper* approach and the *filter* approach [29]. The wrapper approach evaluates different feature subsets' usefulness testing them on a specific learning algorithm. In a typical wrapper approach, a feature subset is selected via a search algorithm and learning models are generated with selected features by applying a cross validation, then accuracies of these models are measured. After these operations, a new feature subset is selected by the search algorithm and previous steps are repeated until the search algorithm terminates. Finally, features that have been used to generate most accurate model are treated as selected features.

Different search algorithms can be used to traverse feature space including, exhaustive search, greedy forward/backward search or genetic algorithms. Although exhaustive search can find the optimum feature subset, other algorithms are generally preferred because multiple models must be generated within a cross validation for each selected feature subset, and this takes a large amount of time. If the number of features is very high, this approach would be infeasible regardless of the selected search method. On the other hand, this approach is multivariate which means presence of different features together is considered. This may be important because some variables that are useless individually may be useful with other features [12].

In the filter approach, features are selected by performing some statistical analysis on features and labels. This approach doesn't require generating learning models as in wrappers and thus filters work faster than wrappers in general. Although it is possible to filter features either individually (univariate[6]) or in subsets (multivariate), most of the filters are univariate. Additionally, selected features by filters can be used on different learning algorithms.

In addition to the filter and the wrapper approaches, some authors mention another method, the *embedded* approach. In this approach, selection operation is done at the training phase by an internal function of learning algorithm (e.g. decision trees). This approach may seem like the wrapper approach at first glance since selection operation is done on a specific learning algorithm. In fact, they are different because embedded approach doesn't include cross validation and repeated learning model generation for different feature subsets as in wrappers.

4 DETECTION ARCHITECTURE

In this section, we present detection architecture in order to make predictions about maliciousness of applications with high accuracy and high sensitivity. As discussed in previous sections, instead of evaluating one aspect of applications with one algorithm, considering different aspects with multiple algorithms may be useful. Discussed architecture has been designed to serve such a purpose. Discussed architecture is component based and each component represents a different type of evaluation in the problem. For example, feature extraction/preprocessing/selection and base learner generation/selection/combination operations can be done using different methods so that each instance of these methods are implemented in different components. For example, static API call and static opcode are two instances of feature extraction components.

There is a dependency among components of discussed architecture. Outputs of some components may be inputs of other components. For example, base learner generation components uses outputs of feature selection components and provides inputs for base learner selection components. Steps to create an ensemble system for proposed architecture are presented in Figure 1. Each step represents a component of the system and dependencies of components can be observed from this Figure.

[6] Also known as feature ranking or weighting.

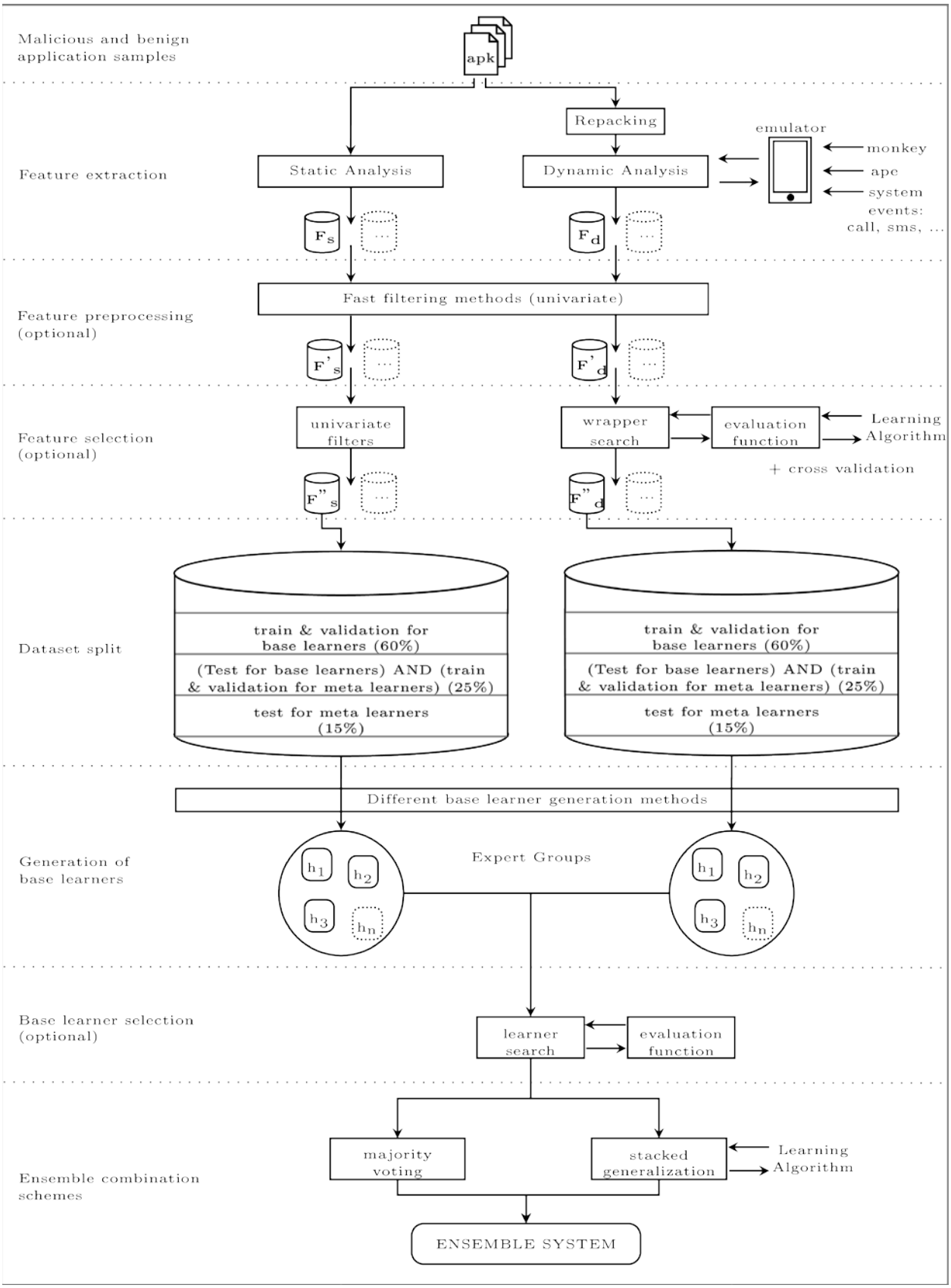


Figure 1: Android malware detection architecture.

As seen in Figure 1, it is possible to extract features via static and dynamic analysis. Although applications can be analyzed directly in static analysis, some preparation may be required before dynamic analysis. In this study, API calls of byte code were collected by repacking and hooking the API calls in byte code. API calls of native code were collected by using a trick on Linux library loader (i.e. LD_PRELOAD trick).

Dynamic analysis is made on an emulator by sending random events via *Android monkey* and also a tool is developed (*ape)* which sends fixed predetermined events such as touch and key events. Additionally, some system events are sent like incoming call, sms and geo location change events during dynamic analysis to reveal malicious intents.

An important point to consider for feature extraction process in Figure 1 is presence of multiple datasets. Assume that, an application set is being analyzed with two static feature extraction components: *native API calls* and *Dalvik Byte API calls*. Number of samples in outputs of these two components may not be same because all Android applications have to contain Dalvik byte code but native code is optional. Hence, sample count for native API call will be smaller. Similar situation is valid for dynamic analysis. It can't be possible to repack and run all applications on emulator, or even if they can be run, enough information may not be gathered.

If the number of extracted features per application is very high, a preprocessing operation can be used to decrease feature size to an acceptable value. Although only univariate filters are displayed in Figure 1 for preprocessing step, it is also possible to use multivariate filters.

Feature selection operations may be used to select most informative features or to increase efficiency and effectiveness.

Feature extraction and selection operations prepare the datasets. In order to build learning models from these datasets, samples for training and testing learners must be specified. For this purpose, datasets are divided into three parts. First part is used to train and validate base learners (60%). Second part is used to test base learners (25%). It is also used for training/validating the meta learners when stacked generalization is used as combination scheme. And third part is used to test meta learners (15%) which are generated using second part's data. If the majority voting is chosen as combination scheme, only third part is used for testing.

Base learners are also known as *experts* or *weak learners*. However, they can be as strong as they can. For example, in order improve generalization, a cross validation is often applied and it is a kind of ensemble method. So that weak learners aren't actually weak. Also, it is possible to use another ensemble learning algorithm (e.g. adaboost) to create base learners as well.

There can be different datasets in the system as discussed previously. Base learners which are generated using the same dataset are called in the same expert group. For instance, base learners which are generated using the dynamic native API call dataset are called as *native API call experts* and their group is called as *native API call expert group*.

Selection of base learners is an optional operation and all base learners may be chosen directly instead of selecting some of them. However, if there are too many learners in the system, selection may improve time efficiency, accuracy etc. Depending on number of learners, different search algorithms and evaluation functions can be used.

Finally, in order to construct ensemble systems, selected base learners are combined. Only majority voting and stacking displayed in Figure 1 but other techniques can be used too.

5 EXPERIMENTAL RESULTS AND EVALUATION

Proposed solution in this study is based on supervised learning. In general, a dataset D is given, which contains samples of the form $D = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)\}$ where $\mathbf{x}_\mathrm{i}$ values are vectors like $< x_1, x_2, \dots, x_n >$, and y_i values are the set of possible labels $y = \{benign, malicious\}$. A hypothesis [7] $h: \boldsymbol{x} \rightarrow y$ is generated using a learning algorithm L and the samples in D. Then, generated hypothesis is used to identify new applications whether they conform to a specific class.

In order to construct an ensemble system, a set of hypotheses H will be generated using a set of datasets $\boldsymbol{D}$ and a set of learning algorithms L. Steps to create these sets are described in the following sections.

5.1 Data Representation

1225 malware samples were used in this study which was collected by Zhou and Jiang [31]. Additionally, 1225 popular applications from different categories were downloaded from *Google Play* via *Google Play Crawler* [5] and these applications considered as benign.

In order to provide a complete evaluation, 4 different feature type were extracted from applications which are believed to complement of each other: Static and Dynamic Native API calls, Static and Dynamic Dalvik Byte API calls.

Static Native API calls to Linux glibc and Android functions were extracted from applications which contain one or more shared objects. Firstly, a vocabulary of used native functions was created ($x = \{log, send, \dots\}$) and then presence information of these functions per application was recorded as a vector like $x = < 0, 1, \dots >$. *Static Dalvik Byte API calls* to Java Core and Android methods were also extracted similar way.

Dynamic analysis produces a sequence of API calls like *log, log, send,....* In order to create a vocabulary from sequence of calls, 2-gram representation was used, whose sliding window was incremented by 1 ($x = \{log + log, log + send, \dots\}$). Presence information of 2-gram sequences was stored in vectors as in static analysis.

[7] I.e. classifier or learning model.

Although most of the methods for Dalvik Byte code were hooked, only a limited native function was hooked for dynamic analysis because hooking all native functions brought an enormous burden to the system since these calls are watched for a system process named *Zygote*[8]. Hence, a limited native functions were chosen by considering static analysis results (write, getuid, getenv, dup, etc.).

As discussed before, the number of samples in different datasets may not be same (e.g. each application doesn't have to contain a native code). Sample count of each created database and the number of features within these datasets is presented in Figure 2. This figure gives a summary of feature extraction, feature filtering, feature selecting and base learner generation steps of proposed architecture. For instance, after the static native (sn) analysis, a dataset was produced which contains 988 samples and 1.147 features. These features were processed with two filtering algorithms and feature size was reduced to 1000. After that, feature selection operations were applied to the filtered features. Information gain and chi square filters were also used as selection algorithms and top 50 features were selected. Additionally, a wrapper approach was applied for two filtered datasets, using forward search as search algorithm and CART as learning algorithm. 3 and 4 features were selected by the wrapper approach based on the error rates for information gain and chi square datasets respectively. Finally, a set of static native dataset was created $D = \{D_{SN1}, D_{SN2}, D_{SN3}, D_{SN4}\}$.

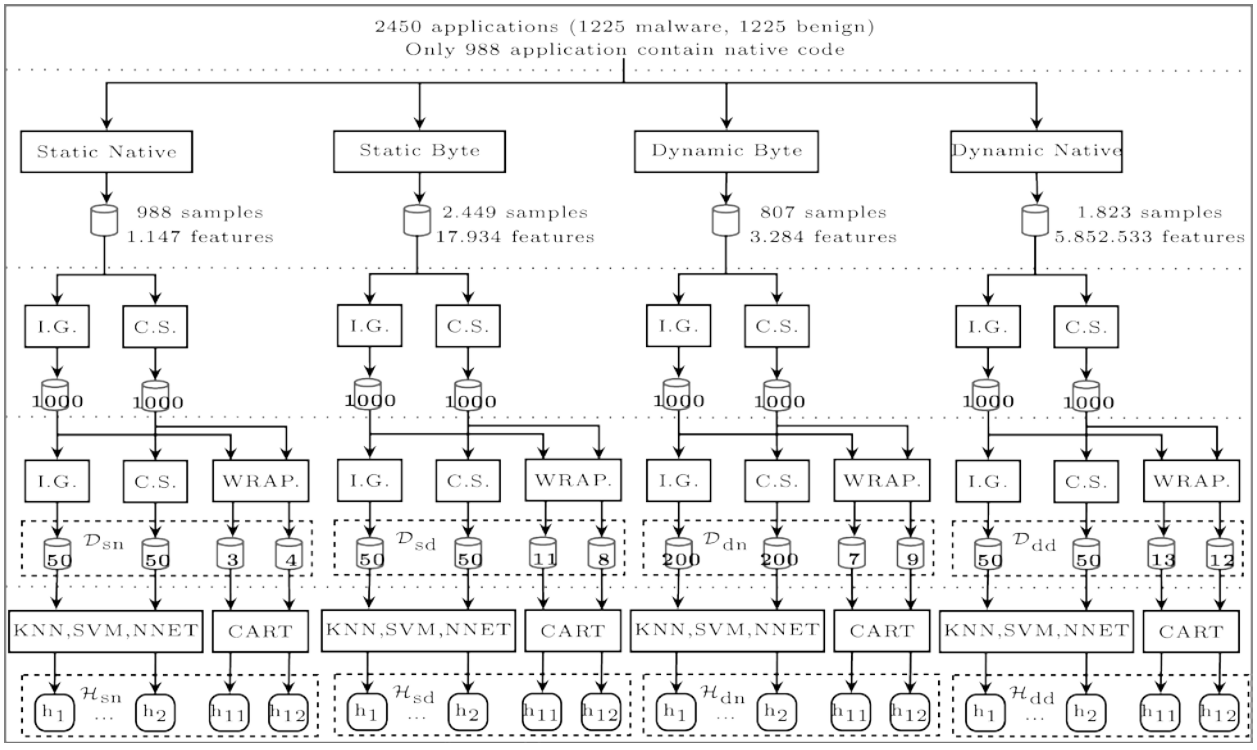


Figure 2: Data representation and base learner generation overview.

[8] Zygote manages all system applications by forking child processes for each one.

Similar steps were applied for the other analysis types. Although most of the steps are identical, number of selected features in dynamic dalvik (dd) analysis was 200 because when 50 features were selected divergence of learning models was very low. At the end, a set of datasets was created which contains all type of dataset sets $\boldsymbol{D} = \{D_{sn}, D_{sd}, D_{dn}, D_{dd}\}$.

5.2 Generating Base Learners

Base learners were generated using the learning algorithms k-NN, NNet, SVMLinear, SVMPoly, SVMRadial and CART. k-NN, NNet and SVM algorithms were used to generate base learners from datasets which were created via filters. On the other hand, CART algorithm was used for datasets that were created using wrapper approach (Fig. 2). General steps for creating base learners are given in Algorithm 1.

Algorithm 1: Steps to generate base learners.

Require:
Datasets for each analysis type: $\boldsymbol{\mathcal{D}} = \{\mathcal{D}_{sn}, \mathcal{D}_{sd}, \mathcal{D}_{dn}, \mathcal{D}_{dd}\}$
Subsets of each dataset: $\mathcal{D}_i = \{D_{i_1}, D_{i_2}, D_{i_3}, D_{i_4}\}$ where $i = \{sn, sd, dn, dd\}$
Learning algorithms: $\mathcal{L} = \{L_{knn}, L_{nnet}, L_{svm}, L_{cart}\}$
Ensure:
A set of all hypothesis: $\boldsymbol{\mathcal{H}} = \{\mathcal{H}_{sn}, \mathcal{H}_{sd}, \mathcal{H}_{dn}, \mathcal{H}_{dd}\}$
Hypothesis per analysis types: $\mathcal{H}_i = \{H_{i_1}, H_{i_2}, H_{i_3}, H_{i_4}\}$ where $i = \{sn, sd, dn, dd\}$

```
1: H ← {}
2: for all D ∈ 𝒟 do
3:     ℋ ← {}
4:     for all D ∈ 𝒟 do
5:         H ← GenerateHypothesis(D, L, 5, 5)
6:         Add H to ℋ
7:     end for
8:     Add ℋ to H
9: end for
   where
1: procedure GenerateHypothesis(D, L, k=5, tune=5)
2:     D_train, D_test ← Split(D) # Split data set to train and test sets
3:     H ← Train(D_train, L, k = 5, tune = 5) # Train a hypothesis
4:     ACC, SNS ← Predict(H, D_test) # Store ACC and SNS from prediction
5:     return H, ACC, SNS
6: end procedure
```

An important point in Algorithm 1 is the parameters of *Train* function. "k" parameter is used in k-fold cross validation. And, "tune" parameter is used to try different meta parameter values for the learning algorithm L. For example, default k value for k-NN is 5 and when this algorithm is used with a tune parameter value 5, $k = \{5, 7, 9, 11, 13\}$ values will be tried for each fold.

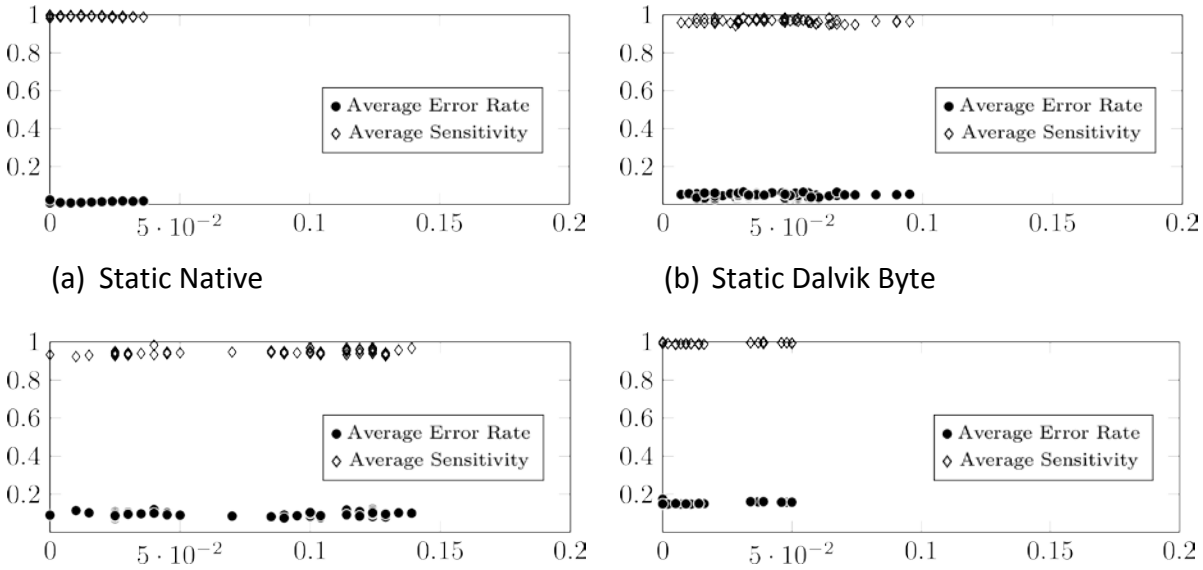
After generating base learners, diversities between each base learner pair were calculated using *disagreement measure* with the formula in Equation 1 assuming K and M are two classifiers, and *a* is the number of samples labeled as malicious by K while they were labeled as benign by M, and *b* is the number of samples labeled as benign by K while they were labeled as malicious by M, *S* is the total sample count. Disagreement measure values which close to 0 means less divergent pairs.

$$dis_{KM} = \frac{a+b}{S} \quad (1)$$

Disagreement measure is a pair wise measurement and it is possible to display produced diversities between base learner pairs. Figure 3 presents diversity-error rate and diversity-sensitivity diagrams using the Equations 2 and 3 for error rate and sensitivity respectively. This figure shows that diverse base learner generation approaches were successful (e.g. using different learning algorithms). For example, there are some base learner pairs whose diversity measure is almost 0.1 in the static Dalvik byte code dataset. Considering the number of samples in this dataset (2.449), diversity measurement value 0.1 of a base learner pair means that pairs have 244 different predictions.

$$Error\ Rate = \frac{\mathrm{FP} + \mathrm{FN}}{TP + FP + TN + FN} \quad (2)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$



(a) Static Native (b) Static Dalvik Byte

(c) Dynamic Native (d) Dynamic Dalvik Byte

Figure 3: Diversity-Error Rate and Diversity-Sensitivity diagrams for each type of dataset. Error rate and sensitivity were denoted by y-axis while diversity was denoted by x-axis.

5.3 Constructing Ensemble Systems

In order to construct ensemble system, three types of combination scheme were applied. First one was a simple majority voting. When number of votes was equal for the majority voting, the final output was produced as *malicious* in order to increase sensitivity. Second scheme was stacking which evaluates all base learners' outputs with multiple learning algorithms. And the third scheme was stacking too but instead of using all base learners, a subset of them was selected before combination using a simple heuristic as in Algorithm 2.

Algorithm 2 proposes a base learner selection approach based on selecting most and least learners considering their accuracy and sensitivity. Since predictions of selected experts will be used to build meta learning models (i.e. stacking), learning the least cases as the most cases would be useful. Hence, base learner pairs were selected using the *FindMost*Pair* and *FindLeast*Pair* methods considering the produced values in Figure 3. When there are multiple pairs with the same accuracy, the most accurate pair was selected by considering the most divergent and the most sensitive pair in order.

Algorithm 2: Base learner selection algorithm.

Require:
A set of all hypothesis: $\mathcal{H} = \{\mathcal{H}_{sn}, \mathcal{H}_{sd}, \mathcal{H}_{dn}, \mathcal{H}_{dd}\}$
Hypothesis per analysis types: $\mathcal{H}_i = \{H_{i_1}, H_{i_2}, H_{i_3}, H_{i_4}\}$ where $i = \{sn, sd, dn, dd\}$
Ensure:
A subset of selected learners: $\mathcal{H}' = \{\mathcal{H}'_{sn}, \mathcal{H}'_{sd}, \mathcal{H}'_{dn}, \mathcal{H}'_{dd}\}$

$\mathcal{H}' \leftarrow \{\}$
for all $\mathcal{H} \in \mathcal{H}$ **do**
  $\mathcal{H}' \leftarrow \{\}$
  $H_{acc} \leftarrow$ FindMostAccuratePair($\mathcal{H}$)
  $H_{l_{acc}} \leftarrow$ FindLeastAccuratePair($\mathcal{H}$)
  $H_{sns} \leftarrow$ FindMostSensitivePair($\mathcal{H}$)
  $H_{l_{sns}} \leftarrow$ FindLeastSensitivePair($\mathcal{H}$)
  Add H_{acc}, $H_{l_{acc}}$, H_{sns} and $H_{l_{sns}}$ to $\mathcal{H}'$
  Add H' to $\mathcal{H}'$
end for

Accuracy and sensitivity results are given in Table 1 for discussed three schemes. As seen, majority voting produced 100% sensitive output for both cases. However, its accuracy was not high as expected. Although selecting a base learner subset hasn't an obvious gain over using all base learners, it can be seen that the two most accurate (97.87%) and the two most sensitive (99.46%) outputs for stacking were obtained from the selected learners.

In Table 1, only accuracy and sensitivity values were given since the motivation of this study is to increase these metrics for malware detection problem. Other metrics can be inference intuitively from these values (e.g. specifity).

Table 1: Accuracy and sensitivity results of combination schemes

	Use All Base Learners		Use Selected Base Learners	
	Accuracy	**Sensitivity**	**Accuracy**	**Sensitivity**
***k*-NN**	97.33	98.91	97.07	98.91
NNet	97.33	98.91	**97.87**	99.46
SVMLinear	94.67	93.48	93.60	93.48
SVMPoly	96.80	98.91	97.07	**99.46**
SVMRadial	97.07	98.37	**97.87**	97.83
CART	96.00	97.28	96. 00	97.28
Majority Voting	91.47	100.0	90.40	100.0

Combining base learners' outputs for stacking requires a missing value handling operation because there will be extra samples in some base learners' outputs since different type of datasets were used. For instance, it isn't sensible to force a native code expert to produce a prediction about an application that doesn't have native code. Hence, we have applied a global constant replacement policy for such missing values to indicate that this expert doesn't have an opinion about related sample. "0" and "1" values were used to indicate *benign* and *malicious* predictions respectively, and "2" was used to indicate *no opinion*.

5.4 Evaluation

As seen in Figure 3, static native analysis gives very accurate and sensitive results. However, only less than half of the applications contain native code. Hence, main point for comparison must be Dalvik byte analysis. Static native Dalvik analysis' best accuracy was 97.22% and best sensitivity was 98.67% which are produced from different base learners. On the other hand, dynamic Dalvik analysis' best sensitivity was 100.0% but best accuracy was only 85.35%.

Ensemble system with the proposed selection algorithm was produced 97.87% accuracy and 99.46% sensitivity. This value is higher than static Dalvik, dynamic Dalvik and dynamic native analyses and also comparable with static native analysis. Furthermore, ensemble system provides more complete solution since different aspects was considered.

As seen from the evaluation results, ensemble learning provided more accurate and complete solution by considering different aspects of applications. On the other hand, since ensemble systems are constructed with lots of base learner, time needed to construct an ensemble system is larger than generating a single learner. However, once features are extracted and selected, base learners can be generated in parallel.

Feature extraction process for dynamic analyses generally takes much longer time than static analyses. In this study, static features were extracted within minutes but dynamic feature extraction took about a week using a single emulator. However, using more than one emulator or device can easily solve this problem.

6 CONCLUSION AND FUTURE WORK

The objective of this study was to solve malware detection problem by proposing architecture based on ensemble learning. For this purpose, different types of features were extracted with multiple methods and these features were processed by multiple mining algorithms in order to generate diverse base learners. Then, results of these base learners were combined in the scope of ensemble learning. Results show that using such architecture increases accuracy and sensitivity of detection operation.

Researchers have proposed lots of Android malware detection studies so far. However, a direct comparison of our results with these studies wasn't supplied for several reasons. First of

all, one of the objectives of this study was to show benefits of ensemble learning for malware detection. Secondly, this study isn't a competitor of other detection tools; some tools can be adapted to this study as a feature extraction component so that they can be used to create new expert groups (e.g. [9]).

Since this study showed benefits of ensemble learning for malware detection problem, we will continue to add and try new components. For example, different evaluation criteria can be used in mining algorithms to improve accuracy and sensitivity. Or, different diversity measurements can be tried. We are also studying to improve accuracy by selecting a subset of benign applications among half a million applications.

7 ACKNOWLEDGMENT

We collected benign applications using "Google Play Crawler", which is developed by Ali Demiroz to support this study, thanks Ali. We also thanks to Zhou and Jiang for sharing their malware collection with us.

"apktool", "smali" and "baksmali" applications were used as reverse engineering tools. Additionally, "R" and its various packages were used for data mining algorithms.

REFERENCES

[1]. Alpaydin, E.: Introduction to Machine Learning (Adaptive Computation and Ma-chine Learning). The MIT Press (2004)

[2]. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. ARTIFICIAL INTELLIGENCE 97, 245-271 (1997)

[3]. Brown, G., Wyatt, J., Harris, R., Yao, X.: Diversity creation methods: A survey and categorisation. Journal of Information Fusion 6, 5-20 (2005)

[4]. Burguera, I., Zurutuza, U., Nadjm-Tehrani, S.: Crowdroid: Behavior-based mal-ware detection system for android. In: Proceedings of the 1st ACM Workshop on Security and Privacy in Smartphones and Mobile Devices. pp. 15-26. SPSM '11, ACM, New York, NY, USA (2011)

[5]. Demiroz, A.: Google Play Crawler (2013), https://github.com/Akdeniz/ google-play-crawler, [Online; accessed 1-April-2014]

[6]. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research 2, 263-286 (1995)

[7]. Dietterich, T.: Ensemble methods in machine learning. In: Multiple Classifier Sys-tems, Lecture Notes in Computer Science, vol. 1857, pp. 1{15. Springer Berlin Heidelberg (2000)

[8]. Deroski, S., enko, B.: Is combining classifiers with stacking better than selecting the best one? Machine Learning 54(3), 255-273 (2004)

[9]. Enck, W., Gilbert, P., Chun, B.G., Cox, L.P., Jung, J., McDaniel, P., Sheth, A.N.: Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. In: Proceedings of the 9th USENIX conference on Operating systems design and implementation. pp. 1-6. OSDI'10, USENIX Association, Berkeley, CA, USA (2010

[10]. Enck, W., Ongtang, M., McDaniel, P.: On lightweight mobile phone application certification. In: Proceedings of the 16th ACM Conference on Computer and Com-munications Security. pp. 235-245. CCS '09, ACM, New York, NY, USA (2009)

[11]. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Annals of Statistics 28, 2000 (1998)

[12]. Guyon, I.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157-1182 (2003)

[13]. Hansen, L., Salamon, P.: Neural network ensembles. Pattern Analysis and Machine Intelligence, IEEE Transactions on 12(10), 993-1001 (Oct 1990)

[14]. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20(8), 832-844 (Aug 1998)

[15]. Kantardzic, M.: Data Mining: Concepts, Models, Methods and Algorithms. John Wiley & Sons, Inc., New York, NY, USA (2002)

[16]. Kuncheva, L., Whitaker, C.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning 51(2), 181-207 (2003)

[17]. Kwok, S.W., Carter, C.: Multiple decision trees. In: Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence. pp. 327-338. UAI '88, North-Holland Publishing Co., Amsterdam, The Netherlands, The Netherlands (1990)

[18]. Partridge, D., Yates, W.B.: Engineering multiversion neural-net systems. NEURAL COMPUTATION 8, 869-893 (1995)

[19]. Petrakos, M., Benediktsson, J.A., Kanellopoulos, I.: The effect of classifier agree-ment on the accuracy of the combined classifier in decision level fusion. IEEE T. Geoscience and Remote Sensing 39(11), 2539-2546 (2001)

[20]. Sahs, J., Khan, L.: A machine learning approach to android malware detection. In: Intelligence and Security Informatics Conference (EISIC), 2012 European. pp. 141-147 (Aug 2012)

[21]. dos Santos, E., Sabourin, R., Maupin, P.: Single and multi-objective genetic al-gorithms for the selection of ensemble of classifiers. In: Neural Networks, 2006. IJCNN '06. International Joint Conference on. pp. 3070-3077 (2006)

[22]. Schmidt, A.D., Schmidt, H.G., Clausen, J., Yksel, K.A., Kiraz, O., Camtepe, A., Albayrak, S.: Enhancing security of linux-based android devices. In: in Proceedings of 15th International Linux Kongress. Lehmann (Oct 2008)

[23]. Shabtai, A., Kanonov, U., Elovici, Y., Glezer, C., Weiss, Y.: andromaly: a behavioral malware detection framework for android devices. Journal of Intelligent Information Systems 38(1), 161-190 (2012)

[24]. Sharkey, A.J., Sharkey, N.E.: Combining diverse neural nets. THE KNOWLEDGE ENGINEERING REVIEW 12, 231-247 (1997)

[25]. Tang, E., Suganthan, P., Yao, X.: An analysis of diversity measures. Machine Learning 65(1), 247-271 (2006)

[26]. Teu, P., Kraxberger, S., Orthacker, C., Lackner, G., Gissing, M., Marsalek, A., Leibetseder, J., Prevenhueber, O.: Android market analysis with activation pat-terns. In: Prasad, R., Farkas, K., Schmidt, A., Lioy, A., Russello, G., Luccio, F. (eds.) Security and Privacy in Mobile Information and Communication Sys-tems, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 94, pp. 1-12. Springer Berlin Heidelberg (2012)

[27]. Tulyakov, S., Jaeger, S., Govindaraju, V., Doermann, D.: Review of classifier com-bination methods. In: In Machine Learning in Document Analysis and Recognition. Informatica 34 (2010) 111118 S. Vemulapalli et al (2008)

[28]. Wolpert, D.H.: Stacked generalization. Neural Networks 5, 241-259 (1992)

[29]. Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. pp. 856-863 (2003)

[30]. Zhou, W., Zhou, Y., Jiang, X., Ning, P.: Detecting repackaged smartphone applica-tions in third-party android marketplaces. In: Proceedings of the second ACM con-ference on Data and Application Security and Privacy. pp. 317-326. ACM (2012)

[31]. Zhou, Y., Jiang, X.: Dissecting android malware: Characterization and evolution. In: Security and Privacy (SP), 2012 IEEE Symposium on. pp. 95-109 (May 2012)

[32]. Zhou, Y., Wang, Z., Zhou, W., Jiang, X.: Hey, you, get off of my market: Detecting malicious apps in official and alternative android markets. Proceedings of the 19th Annual Network and Distributed System Security Symposium pp. 5-8 (2012)

[33]. Zhou, Z.H., Wu, J., Tang, W.: Ensembling neural networks: Many could be better than all. Artificial Intelligence 137(12), 239-263 (2002)

TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 2, ISSUE 3
ISSN: 2054 - 7390

Estimation of solar radiation power using reference evaluation of solar transmittance, 2 bands (REST 2) model

(Case study : Semarang, central java, Indonesia)

Benedictus Asriparusa
Institut Teknologi Bandung

ABSTRACT

Indonesia is a country which has abundant energy resources of solar demonstrated by potential position of solar annual migration (around equator line). Nowadays fossil energy consumption so apprehensively, these things are due to decreasing of fossil energy while demand of fossil energy rises continuously. In 2011, fossil fuel energy of Indonesia is 66.4% besides negative effect of increasing Greenhouse Gases concentrate, which is increasing of surface temperature and creates inconvenience environment. Sun as the biggest energy resource should be use optimally for Indonesia area. Diversification of energy is a final step to get another resource so release us of dependently fossil resources. Solar radiation is a green renewable energy which has the potential to answer the needs of energy problems on the period. Knowing how to estimate the strength of the solar radiation force may be one solution of sustainable energy development in an integrated manner. Unfortunately, a fairly extensive area of Indonesia is still very low availability of solar radiation data. Therefore, we need a method to estimate the exact strength of solar radiation. In this study, author used a model Reference Evaluation of Solar Transmittance, 2 Bands (REST 2). Validation of REST 2 model has been performed in Spain, India, Colorado, Saudi Arabia, and several other areas. But it is not widely used in Indonesia. Indonesian region study area is represented by the area of Semarang, Central Java. Solar radiation values estimated using REST 2 model was then verified by field data and gives average RMSE value of 6.53%. Based on the value, it can be concluded that the model REST 2 can be used to estimate the value of solar radiation in clear sky conditions in parts of Indonesia.

Keywords : Estimation, Solar Radiation Power, REST 2, Renewable Energy

1 INTRODUCTION

Energy currently holds an important role in the development of economic sector in National Country. It would be a thing which is not disputed and often regarded as the most powerful of

DOI: 10.14738/tmlai.23.210
Publication Date: 9th June 2014
URL: http://dx.doi.org/10.14738/tmlai.23.210

the economy. This thing is recognized by developing countries in the world. Then, they always think that how to implement energy usage accurately and developing other technologies efficiently as an main requirement to increase economic sector. Indonesia is one of the developing countries that has different types of energy resources in quality and quantity. Generally, Indonesia could manage of energy resources appropriately to improve social welfare in that country.

Indonesia's location is around the equator line, which is the latitude 6^0 NL – 11^0 SL and the longitude 95^0 EL - 141^0 WL. Based on the circulation of the sun in a year (which is at 23.5^0 NL and 23.5^0 SL area), Indonesian territory will be exposed by the sun during 10-12 hours per day. Due to the location of Indonesia and position of the sun, that's why Indonesia has a solar radiation level is very high. According to the measurement from the center of Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) Indonesia, they estimated that the solar radiation touching on the surface of the Earth Indonesia (especially Eastern Indonesia) on average ± 5.1 kWh/m^2 per day and given variation around 9% per month. (NN, 1994).

Based on the research, they are trying to simulate the influence of cloud sheet towards on global energy balance shows that albedo (the deviation of solar radiation are reflected and received by the earth) increases continuously from 15% to 30% which means that the quantity of energy loss is 50W/m^2. Clouds also reduce the emission of infrared light until 30 w/m^2. It makes the effect of clouds sheet in global balance system loss the energy until 20 W/m^2. While the quantity of greenhouse effect given 4 W/m^2 of global warming, even though it given the addition of the CO_2 content in the atmosphere is two times larger than the current state. It shows that method of collection Solar radiation data still not effective. (Intergovernmental Panel on Climate Change, 2001).

REST Model 2 is one of the model method to estimate the value of solar radiation for clear-sky Model Condition. This model take into measurement of atmospheric condition like humidity, water vapor, ozone layer, precipitation, and contain of gas that can be reflected solar radiation to the outside of Earth. REST Model 2 also uses the data to calculate aerosol turbidity like CO_2 and NO_2. (Gueymard C. A., 2009).

The focus area is Semarang, Central Java. It is represented by point Climatological Station in Semarang (06.967 SL and 110.417 EL). Time studies are in January and July of 2011. This research aimed to clear-sky conditions. Bandung area as comparison to confirm the validity of the model REST 2. It is represented by the location in 06.8 SL and 107.4 EL with time studies throughout the month of January 2010 and June 2010.

This research has four objectives. Those are to determine the time of maximum and minimum value of solar radiation through on estimation, to simulate the movement of CO_2 and NO_2 in

Semarang Area, to find out whether the REST 2 model is suitable applied in Indonesia., and to looking for other alternatives to fill the solar radiation data other than direct measurement.

Reference Evaluation of Solar Transmittance, 2 bands (REST 2) is a two-band models developed by Christian A. Gueymard in 2008. Band 1 covers UV wave and Visible wave, from 0.29 to 0.70 µm. The characteristic of Band I is huge absorption by ozone in the UV wave and scattering by molecules and aerosols. Band II covers near infrared wave, from 0.7 to 4 µm. The characteristic of Band II is huge absorption by H_2O, CO_2, NO_2, and other gases. Model REST 2 is an updated version of the previous model ie the Reference Evaluation of Solar Transmittance (REST). REST 2 Model is considering absorption by NO_2 (Gueymard CA, 2002) because absorption of NO_2 is an important parameter in estimating the value of the solar radiation. (Rizwan Jamil, & Kothari, 2010).

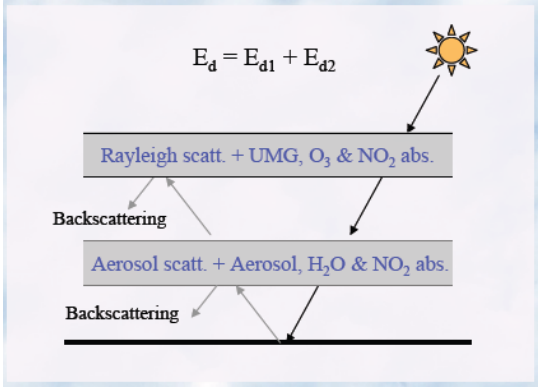


Figure 1 Scattering of two layers Scheme (Gueymard C. A., 2009)

$$E_d = E_{d1} + E_{d2}$$

$$E_{d1(band1)} = f_1 \, R \, E_{SC} \, T_{R1} \, T_{G1} \, T_{O1} \, T_{N1} \, T_{W1} \, T_{A1}$$

$$E_{d2(band2)} = f_2 \, R \, E_{SC} \, T_{R2} \, T_{G2} \, T_{O2} \, T_{N2} \, T_{W2} \, T_{A2}$$

Generally, REST 2 Model as clear-sky model condition has 8 input data. It is base on satellites measurement that is MRM-5 (5data input), CSR (5 input data), METSTAT (5 input data) , Yang (4 input data) , Heliosat-2 (2 input data). It is also equipped with statistical measurement that is mean bias difference (MBD) and root mean square difference (RMSD).

Based on previous research, it shows that value of root mean square (RMSE) reached 5.93% in Saudi Arabia. (Shafiqur, 1997). One of research in India also shows that the value of root mean square (RMSE) reached 3.4 % (Rizwan, et al, 2010), research in Germany shows that the value of root mean square reached 5.1% (Lorenz, 2013) and research in Spain shows that the value of root mean square reached 9% (Mateos et al, 2010). Thus, it means RMSD of REST 2 were in the range 0.7 % to 11.9% (Gueymard CA, 2011)

According to all information, it can conclude that the deviation value for REST 2 model is in the range 0.7% to 11.9%. Therefore, the value of deviation is acceptable to show that the results of

the REST 2 models can be used to estimate the solar radiation in a region. So, author determined that the threshold for the value of RMSE is 11.9%.

2 METHOD OF MODEL EVALUATION

On this research, author is divided into four main parts: input data model base on observation, satellite data, modeling of dominant gas of atmosphere and data correlation to verify the model. There are 8 input data required by the model REST 2: precipitation of water, surface pressure , Angstrom's wave exponent (α), Angstrom turbidity factor, aerosol single-scattering albedo (using default), the total columnar ozone amount, total columnar nitrogen dioxide amount (using default), the solar zenith angle (calculated from the date and time).

Based on observation process, solar radiation data obtained from measurements using the Automatic Weather Station (AWS) with brand VAISALA placed in Badan Meteorologi Klimatologi dan Gofisika Semarang data available with an interval of one second. Data can be downloaded in http://aws.bmkg.go.id/MetView / # dataquery (AWS Center, 2013).

Based on satellite measurement, satellite is used satellite MTSAT. The image used is the visible image of the canal with a wavelength of 0.55 - 0.80μm and a resolution of 1 km. Data can be downloaded via http://weather.is.kochi-u.ac.jp/sat/GAME (Index of/sat/GAME/2011, 2005).

Based on simulation of dominant gas of atmosphere, solar radiation data for Bandung area obtained by recording data that is placed in the Garden of AWS Meteorological ITB. Data distribution of NO_2 and CO_2 using a simulated distribution TAQM (Taiwan Air Quality Model).

The step on this research is to collect input data required to run the model REST 2 according to observation and data measurement. The next step is the preparation of model input. Models are arranged and grouped by month. So, there are 7 input files for each month. We have 4 times per day, then the input set 07.0 pm (00:00 UTC), 10:00 pm (03:00), 13:00 pm (06:00 UTC), and 16:00 (09:00 UTC) as time arranged.

The next step is determined by time case during a *clear-sky* condition. Determine of *clear-sky* condition depends on looking the area without cloud at satellite images. Afterwards, verify the results of the REST 2 model output used 3 methods of statistical tests that is root mean-square error (RMSE), correlation, and the simulated distribution of the dominant atmospheric gases (CO2 and NO2).

Root mean square error shows that :

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{R_{ob}} - \mathrm{R_{mod}})^2}$$

According the results, we obtained that verification of data error gap is quite large. So we have to give some corrections to model in order to provide good results. Data observation is data received on the surface of the earth after undergoing attenuation, while data model is the solar radiation data received by the earth in the clear-sky conditions. Ratio of solar radiation received by the earth's surface is observational data divided by data model. The ratio will be provided empirical formula as a correction to the model results.

3 RESULT AND DISCUSSION

3.1 REST 2 Model Result

3.1.1 Semarang

The blue line is the result of solar radiation model, while the red line is the solar radiation observation. Graphical image shows that the blue line values has trend pattern to be higher than the red line. This is acceptable because radiation of the model is the radiation received by the earth's radiation under clear-sky conditions, while radiation of the observation is actual radiation received by the earth where there are factors that reduce the radiation that is not accounted for by the model. It is caused by cloud.

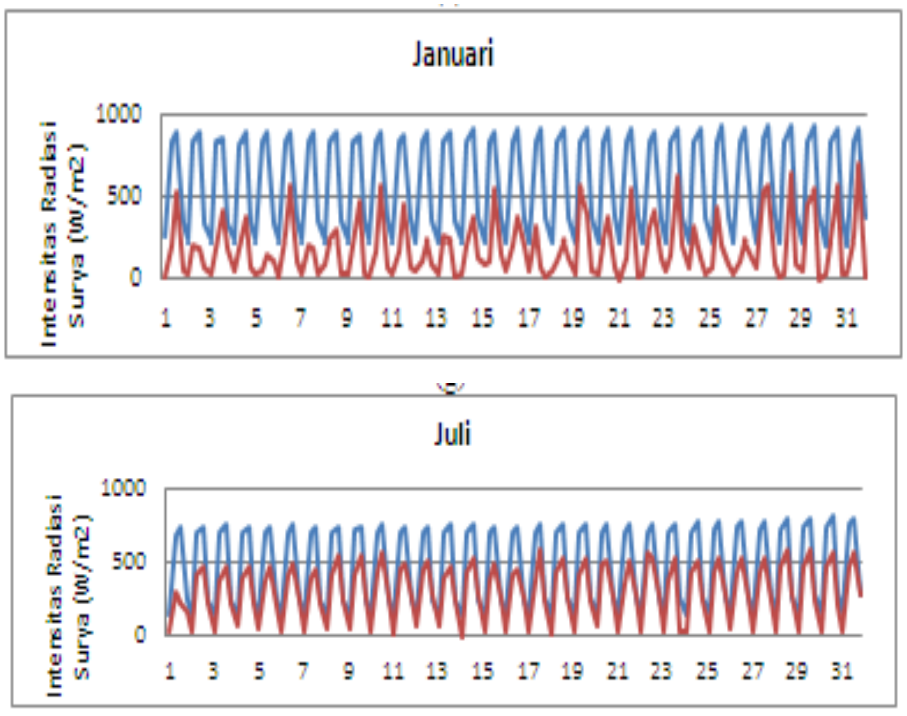


Figure 1 Solar radiation of REST 2 model (Blue line) and solar radiation of observation data (red line)

Generally, The first month (January) is the peak of the rainy season so that the growth of the cloud is very large compared to other months, while June and July are already entering the dry months so that the growth of the cloud tend to be smaller. It gives impact to the solar radiation received by the earth's surface which means that clouds are blocked all radiation that will come to earth's surface.

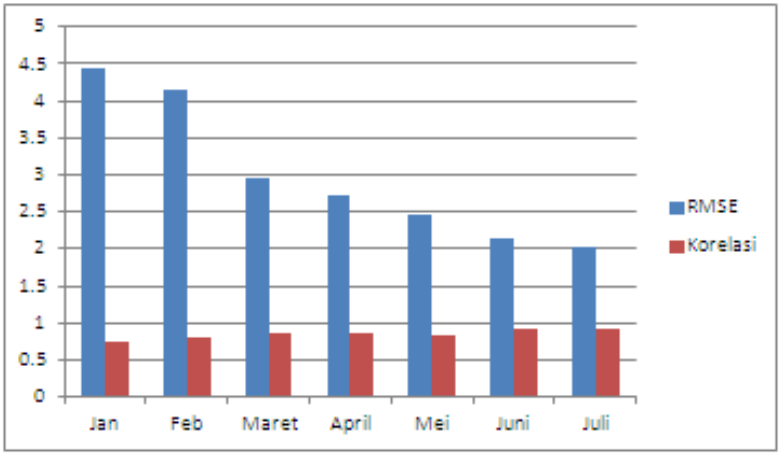


Figure 2 RMSE and correction for each month on percentage

The actual radiation in January will be smaller than in June and July, it can be seen from the RMSE is on above the graphic. RMSE in January reached 444.8W/m2 and RMSE in June and July, respectively 213.4 W/m2 and 203.3 W/m2 without gas dominant measurement, while RMSE in January reached 452.3 W/m2 and RMSE in June and July, respectively 220.6 W/m2 and 211.7 W/m2 with gas dominant measurement.

Solar radiation generated by the model REST 2 is the maximum possible values solar radiation received by the solar radiation of earth's surface. The average value in Semarang area with clear-sky condition and gas dominant of atmosphere is reached to 532.2124 W /m2, while the maximum value is equal to 954.2 W/m2 (February). The maximum value for each month are: January (934 W/m2). February (954.2 W/m2), March (945.2 W/m2), April (884.7 W/m2), Mei (795.1 W/m2), June (760.8 W/m2), and July (815.1 W/m2).

3.1.2 Bandung

In Bandung area, we also applied same method to determine solar radiation maximum and RMSE. According to calculation, it shows that RMSE value in January 2010 reaches 121.23 W/m2 while RMSE value in June 2010 reaches 84.32 W/m2. So, Bandung area as comparison shows that it confirms the REST 2 model can be applied at region of Indonesia.

3.2 Clear-Sky Condition

3.2.1 Semarang

Author have to arrange and determine the times where the research is clear-sky conditions used satelit (MTSAT). The observation of cloud cover shows in 110.417 LN and 06.967 LS point. Image A shows that clear-sky condition at Semarang area, Central Java and image B shows that satellite images which are not included the clear-sky condition.

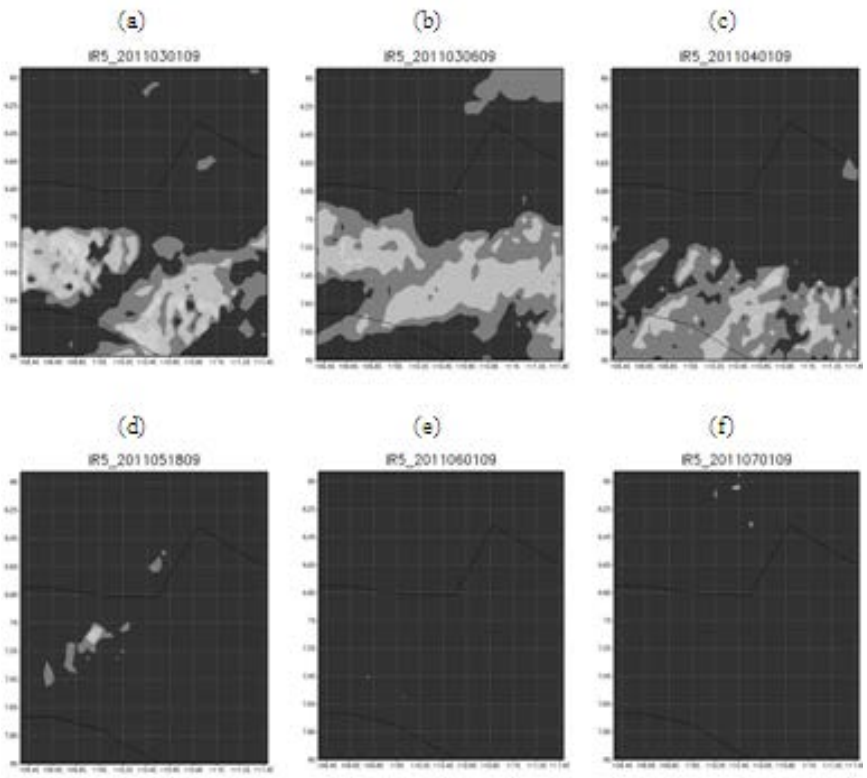
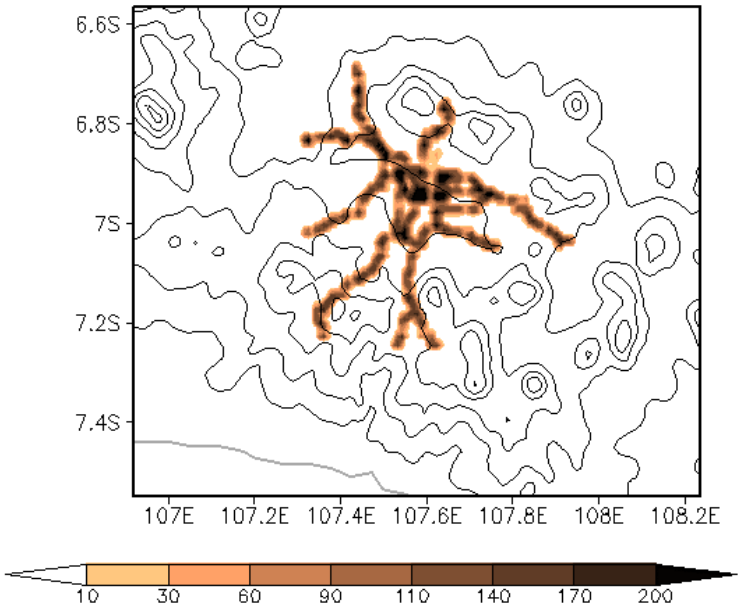


Figure 3 MTSAT satellite images (Satellite images of Central Java with Semarang area (in 110.417 LN and 06.967 LS)

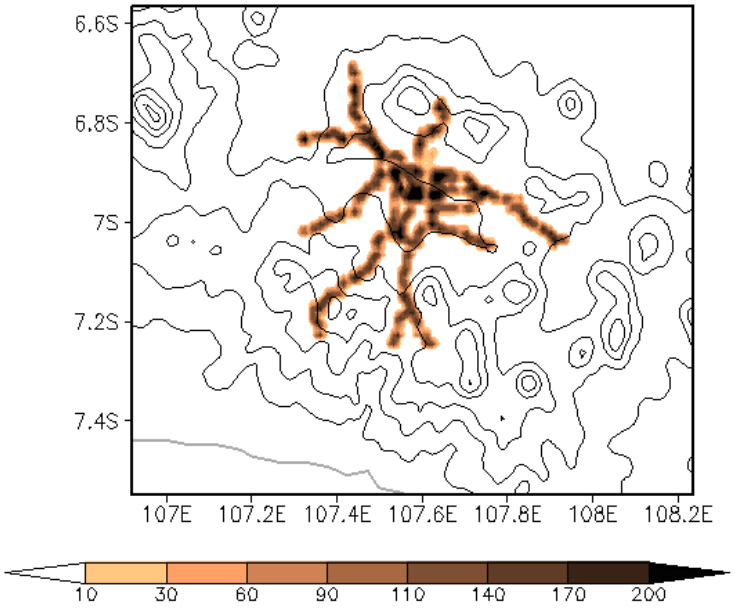
According to MTSAT satellites images, we found 67 times out of a total of 844 times that clear-sky conditions. Clear-sky conditions most commonly found in June and July, respectively 19 times and 27 times. Total RMSE for clear-sky conditions 17.23 W/m2. The average RMSE is equal to 6.53421%, and the value is smaller than the threshold value. RMSE of REST 2 model shows that the model can estimating solar radiation well in clear-sky conditions for the area of Semarang, Central Java.

3.2.2 Bandung

Clear-sky conditions at the Bandung area has quite small RMSE value. It is equal to 14.72979 W/m2 without taking into account the distribution of CO2 and NO2 (Tika, 2012). This is because the correction does not consider the simulation of the spread of water vapor, CO2, NO2, and wind. Below is a picture of the distribution of CO2 in Bandung. Image to the left is a picture of the distribution of CO2 in Bandung area, West Java (6.8 LS and BT 107.4) in January and to the right of the picture is in June.



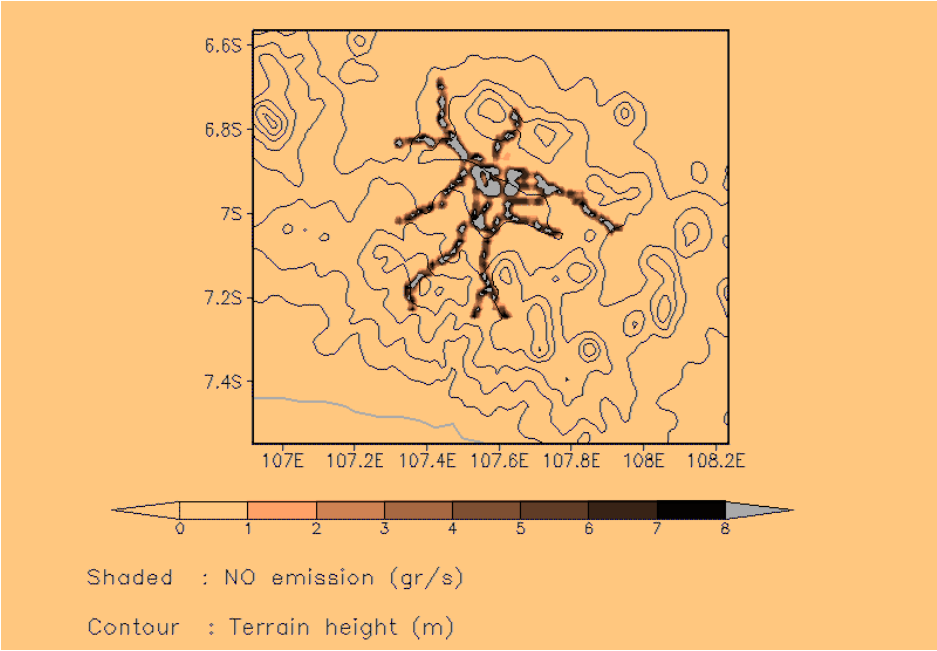
(a)



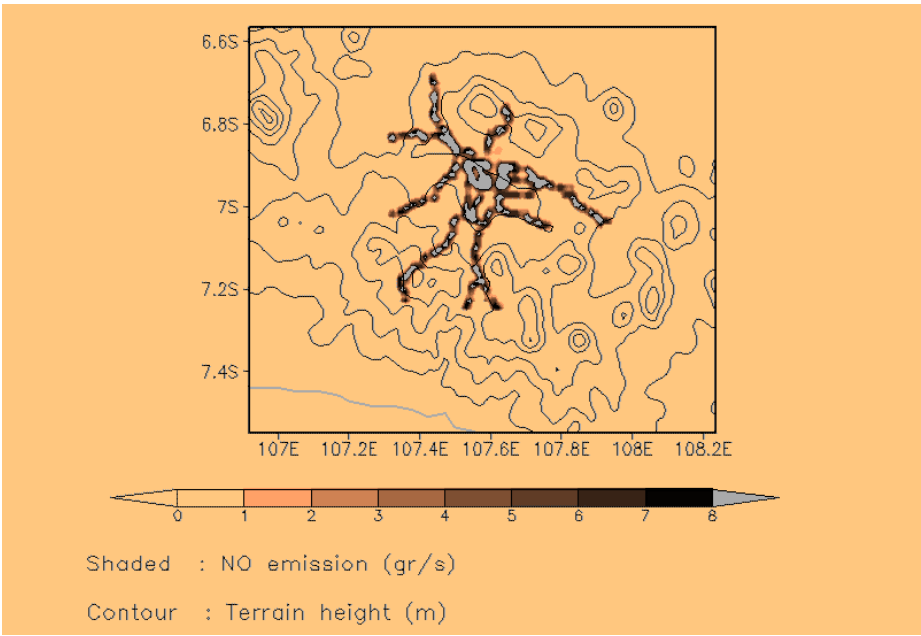
(b)

Figure 4 Simulation of Dissemination CO2 on (a) January and (b) June

RMSE values with CO2 measurement shows that it is equal to 20.36767 W/m2. If we look to the distribution of NO2 in the Bandung area, West Java, we can see the picture to the left of the distribution of NO2 in January and to the left of the picture in June.



(a)



(b)

Figure 5 Simulation of Dissemination NO2 on (a) January and (b) June

RMSE values with distribution of NO2 measurement shows that it is equal to 27.81282 W/m2. This suggests that the distribution of CO2 and NO2 have considerable influence on the development of simulation models and we need to make the distribution of dominant gases in the atmosphere prior to provide better precision.

The value of REST 2 confirms that the model must be modified and verified with the dominant model of the gas distribution in the atmosphere firstly and then we currently can be used it to estimate the solar radiation in clear-sky conditions for Indonesian territory.

3.3 Correction and Evaluation of Model

3.3.1 Semarang

Based on previous result, we can see that the value of RMSE is very large. So it takes corrective measurement to improve the results of the REST 2 model. This correction gives mean that we have to count parameterization factors on the model REST 2, which is cloud factors. First step is we must find the ratio of solar radiation received by the earth's surface. The ratio is obtained by dividing the solar radiation observations to the results of model calculations of solar radiation.

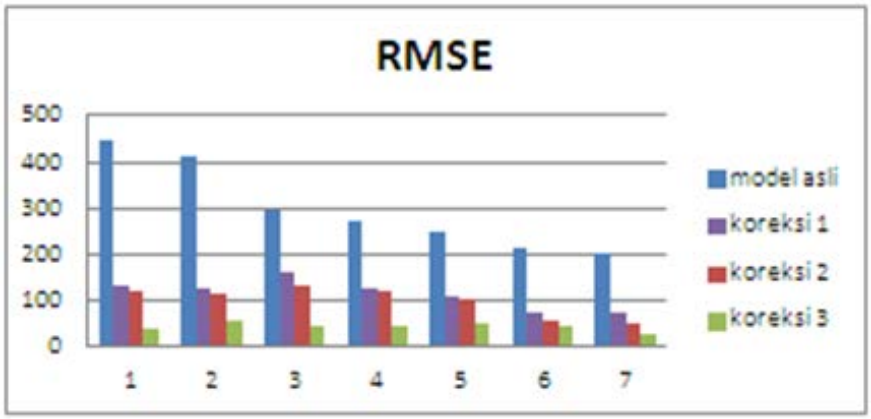


Figure 6 RMSE value comparison

In first experiment correction, ratio will calculate for each time. Then the average of RMSE values over each month has its own tendencies are influenced by season. RMSE in January reached 147.2 W/m2, March reached 179.4 W/m2 (largest RMSE) and in July reached 82.1 W/m2 (smallest RMSE).

In second experiment, we consider the correction due to the influence of the revolution of the earth's seasons and the growth of the cloud due to the movement of the earth's rotation. The growth of cloud between the morning and afternoon will distinguish.

Correlation of second experimental results gives improvement, which shows that the correlation was reduced from 0.79 to 0.77. RMSE was getting better. The large of RMSE amounting to 135.2 W/m2 in March, while the smallest was in July of 54.1 W/m2.

In third correction, the results are getting better. The correlation increased to 0.94, then the smallest correlation was 0.82 in January and the largest is 0.95 in July. RMSE improved

significantly. RMSE average reached 48.2 W/m2, which means shrank 83.4 % from the previous value 290.6 W/m2. (The average of RMSE with original model).

3.3.2 Bandung

Based on previous result, we can see that the value of RMSE is improved significantly, then the correction step should be applied to other areas to ensure that the method can be used with either. The next correction method will try to the area of Bandung.

Once applied correction method, RMSE value of 57.48 W/m2 which is shrinking 32%. So, it means that the correction methods just only to correct the model output, but the results are not very significant.

REST 2 model is indeed a model for clear-sky conditions, but there needs to be a special correction on simulation of distribution dominant gases in the atmosphere when used directly without modification for the region with the cloud growth is high regardless of the dominant gases such as Indonesia, it will not be provide quite representative output.

4 CONCLUSIONS

According to this research, we can put some conclusions. The conclusions are :

- The average of estimation solar radiation value in point of Badan Meteorologi Klimatologi dan Geofisika (BMKG) Semarang is 532.2124 W/m2 (clear-sky condition and gas dominant measurement)
- The maximum value of point Badan Meteorologi Klimatologi dan Geofisika (BMKG) Semarang for each month are: January (934 W/m2). February (954.2 W/m2), March (945.2 W/m2), April (884.7 W/m2), Mei (795.1 W/m2), June (760.8 W/m2), and July (815.1 W/m2).
- Model REST 2 must used by additional data like simulation of dominant gas of atmosphere to get better precition of result so we could use it to prediction maximum value of solar radiation received by earth's surface.
- In clear-sky conditions with the simulation of CO2 and NO2, RMSE total is 17.23 W/m2. If we converted into a percentage, the average RMSE is equal to 6.53421% (6.53%).
- REST Model 2 can be applied in Indonesia, but only for clear-sky conditions and need additional data such as the simulation of the dominant atmospheric gases to obtain more precise results.
- If we will calculate the model in a state of cloudy skies. It means the model cannot be applied directly but requires correction step (the dominant atmospheric gases, water vapor, wind, etc.).
- REST Model 2 can be used as an alternative to predict and estimate the solar radiation data other than direct measurement.

REFERENCES

Sasongko, Tika. (2013). Perhitungan Radiasi Surya menggunakan Reference Evaluation of Solar Transmittance, 2 bands (REST 2) model (Studi kasus: Semarang).

Rahman, Shafiqur. (2007). Solar Radiation Over Saudi Arabia And Comparisons With Empirical Models. Retrieved January 09, 2014, from Jurnal of Saudi Arabia Center.

Lorenz, Elke. (2013). Current status of solar PV Power forecasting. Retrieved January 09, 2014, from Jurnal and Workshop Universitat Oldenburg : http://www.3e.eu/wp-content/uploads/2013/11/Lorenz_MeteoRES-Workshop_2013-11-19.pdf

PSD Climate Data Repository. (2011). Retrieved November 25, 2013, from Earth System Research Laboratory: http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html

AWS Center. (2013). Retrieved November 2013, from http://aws.bmkg.go.id/MetView/#dataquery

Gueymard, C. A. (2002). Direct solar transmittance and irradiance predictions with broadband models. Part I: detailed theoretical performance assessment. *Solar Energy* , 361.

Gueymard, C. A. (2009). Retrieved January 11, 2014, from Solar Consulting Service: http://www.solarconsultingservices.com/ReadMe_REST2_v7.1.txt

Index of /sat/GAME/2011. (2005, 11 17). Retrieved January 15, 2014, from Kochi University: http://weather.is.kochi-u.ac.jp/

Gueymard, C. A. (2011). Clear-sky irradiance predictions for solar resource mapping and large-scale applications: Improved validation methodology and detailed performance analysis of 18 broadband radiative models. *Solar Energy* .

Justus, C. G., & Paris, M. V. (1985). A Model For Solar Spectral Irradiance And Radiance at the Bottom and Top of A Cloudless Atmosphere. *Journal of Climate and Applied Meteorology* .

Mateos, D., Bilbao, J., Miguel, A. d., & Burgos, A. P. (2010). Prediction of Solar Irradiance And Illuminance Using REST2 Model. Valladolid, Spain.

Rahardjo, I., & Fitriana, I. (2006). Analisis Potensi Pembangkit Listrik Tenaga Surya di Indonesia. *Strategi Penyediaan Listrik Nasional Dalam Rangka Mengantisipasi Pemanfaatan PLTU Batubara Skala Kecil, PLTN, Dan Energi Terbarukan* , 43-52.

Rizwan, M., Jamil, M., & Kothari, D. P. (2010). Solar energy estimation using REST2 model. *International Journal of Energy and Environment* .

NOMENCLATURE

RMSE = root-mean-square error

RMSD = root-mean-square deviation

n = total data that will be used

R_{ob} = solar radiation base on observation

R_{mod} = solar radiation base on REST 2 model

R = Rayleigh scatter

G = mixed gas absorption

O = Ozone absorption

N = NO_2 absorption

W = water vapor absorption

A = scatter and absorption of aerosols