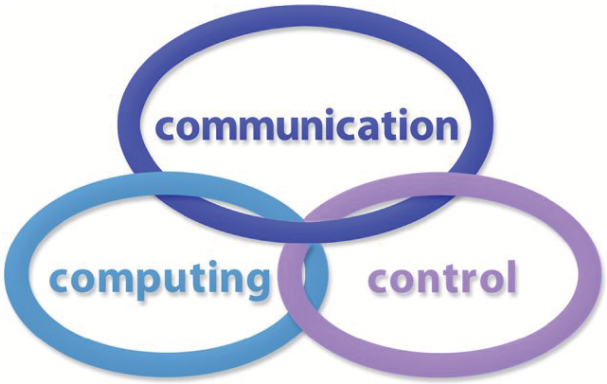
INTERNATIONAL JOURNAL

of

COMPUTERS COMMUNICATIONS & CONTROL

ISSN 1841-9836



A Bimonthly Journal
With Emphasis on the Integration of Three Technologies

Year: 2014 Volume: 9 Issue: 5 (October)

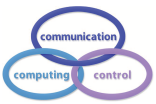
This journal subscribes to the principles of the Committee on Publication Ethics (COPE).

CCC Publications - Agora University Editing House

CCC Publications

http://univagora.ro/jour/index.php/ijccc/

International Journal of Computers Communications & Control



Editorial Address:
Agora University/ R&D Agora Ltd. / S.C. Cercetare Dezvoltare Agora S.R.L.
Piaţa Tineretului 8, Oradea, jud. Bihor, Romania, Zip Code 410526
Tel./ Fax: +40 359101032
E-mail: ijccc@univagora.ro, rd.agora@univagora.ro, ccc.journal@gmail.com
Journal website: http://univagora.ro/jour/index.php/ijccc/

Gheorghe Păun
Institute of Math. of Romanian Academy
Bucharest, PO Box 1-764, Romania
gpaun@us.es

Mario de J. Pérez Jiménez
Dept. of CS and Artificial Intelligence
University of Seville, Sevilla,
Avda. Reina Mercedes s/n, 41012, Spain
marper@us.es

Dana Petcu
Computer Science Department
Western University of Timisoara
V.Parvan 4, 300223 Timisoara, Romania
petcu@info.uvt.ro

Radu Popescu-Zeletin
Fraunhofer Institute for Open
Communication Systems
Technical University Berlin, Germany
rpz@cs.tu-berlin.de

Imre J. Rudas
Óbuda University
Budapest, Hungary
rudas@bmf.hu

Yong Shi
School of Management
Chinese Academy of Sciences
Beijing 100190, China &
University of Nebraska at Omaha
Omaha, NE 68182, U.S.A.
yshi@gucas.ac.cn, yshi@unomaha.edu

Athanasios D. Styliadis
University of Kavala
Institute of Technology
65404 Kavala, Greece
styliadis@teikav.edu.gr

Gheorghe Tecuci
Learning Agents Center
George Mason University
U.S.A.
University Drive 4440, Fairfax VA
tecuci@gmu.edu

Horia-Nicolai Teodorescu
Faculty of Electronics and
Telecommunications
Technical University "Gh. Asachi" Iasi
Iasi, Bd. Carol I 11, 700506, Romania
hteodor@etc.tuiasi.ro

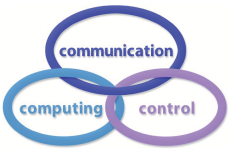
Dan Tufiş
Research Institute for Artificial Intelligence
of the Romanian Academy
Bucharest, "13 Septembrie" 13, 050711,
Romania
tufis@racai.ro

Lotfi A. Zadeh
Director,
Berkeley Initiative in Soft Computing (BISC)
Computer Science Division
University of California Berkeley,
Berkeley, CA 94720-1776
U.S.A.
zadeh@eecs.berkeley.edu

DATA FOR SUBSCRIBERS
Supplier: Cercetare Dezvoltare Agora Srl (Research & Development Agora Ltd.)
Fiscal code: 24747462
Headquarter: Oradea, Piata Tineretului Nr.8, Bihor, Romania, Zip code 410526
Bank: BANCA COMERCIALA FEROVIARA S.A. ORADEA
Bank address: P-ta Unirii Nr. 8, Oradea, Bihor, România
IBAN Account for EURO: RO50BFER248000014038EU01
SWIFT CODE (eq.BIC): BFER

BRIEF DESCRIPTION of IJCCC

Title of journal: International Journal of Computers Communications & Control.
Acronym: IJCCC; **Starting year of IJCCC:** 2006
Abbreviated Journal Title in JCR: INT J COMPUT COMMUN.
International Standard Serial Number: ISSN 1841-9836, ISSN-L 1841-9836
Publisher: CCC Publications - Agora University of Oradea.
Publication frequency: Bimonthly: Issue 1 (February); Issue 2 (April); Issue 3 (June); Issue 4 (August); Issue 5 (October); Issue 6 (December).
Founders of IJCCC: Ioan Dzitac, Florin Gheorghe Filip and Mişu-Jan Manolescu.
Logo:



Coverage:

- Since 2006, Vol. 1 (S), IJCCC is covered by Thomson Reuters and is indexed in ISI Web of Science/Knowledge: Science Citation Index Expanded.
- Journal Citation Reports (JCR - Science Edition), IF = 0.694 (JCR2013).
  Subject Category:
  - Automation & Control Systems: Q4 (46 of 59);
  - Computer Science, Information Systems: Q3 (96 of 135).
- Since 2008, 3(1), IJCCC is covered in Scopus, SJR2013 = 0.231, H index = 13.
  Subject Category:
  - Computational Theory and Mathematics: Q4;
  - Computer Networks and Communications: Q3;
  - Computer Science Applications: Q3.
- Since 2007, 2(1), IJCCC is covered in EBSCO.

Focus & Scope: International Journal of Computers Communications & Control is directed to the international communities of scientific researchers in computer and control from the universities, research units and industry.

To differentiate from other similar journals, the editorial policy of IJCCC encourages the submission of original scientific papers that focus on the integration of the 3 "C" (Computing, Communication, Control).

In particular the following topics are expected to be addressed by authors:

- Integrated solutions in computer-based control and communications;
- Computational intelligence methods (with particular emphasis on fuzzy logic-based methods, ANN, evolutionary computing, collective/swarm intelligence);
- Advanced decision support systems (with particular emphasis on the usage of combined solvers and/or web technologies).

Contents

International Journal of Computers Communications & Control, ISSN 1841-9836
Volume 9, Issue 5, October 2014.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 9(5):531-538, October, 2014.

Estimation of the Technical State of Automotive Disc Brakes Using Fuzzy Logic

M. Baban, C.F. Baban, C. Bungau, G. Dragomir, R.M. Pancu

Marius Baban, Calin Florin Baban*,
Constantin Bungau, George Dragomir, Rares Mihai Pancu
University of Oradea, Romania, Oradea, Universitatii st., 1
mbaban@uoradea.ro, bungau@uoradea.ro,
georgedragomir@yahoo.com, pancurares@yahoo.com
*Corresponding author: cbaban@uoradea.ro

Abstract: According to existing studies the phenomena that occur in the exploitation of the braking system are very complex and an analytical mathematical modeling of braking process it is difficult to be developed [1]. Since these phenomena are also characterized by some uncertainties, a fuzzy logic approach has been employed in this research for the estimation of technical state of the disc brakes. Their technical state was expressed through the thickness variation, which was used as the output linguistic variable. The vibrations and temperature of the disc brakes were used as the input linguistic variables. The fuzzy decision system for the estimation of technical state of the disc brakes has been implemented with the Fuzzy Logic Toolbox ^{TM} of the Matlab ® software, which can be employed to determine if the thickness of the disc brakes becomes smaller than the limit value prescribed by the manufacturer.
Keywords: disc brake, temperature, vibration, thickness variation, fuzzy logic approach.

1 Introduction

Throughout the exploitation of the disc brake its thickness is modified in the contact interface with the friction material of the brake pads, due to the effect of wear. If the thickness of the disc brake becomes smaller than a certain limit value prescribed by the manufacturer, the probability of the deformation of the disc brake or even its damage is increasing at the high intensity or long duration of braking. This may have serious implications for the road accidents.

The replacement of discs and brake pads depends on the number of traveled kilometers and it is usually done preventively, when periodic revisions are carried out. Such replacement can occur even if maximum wear is not reached, which leads to higher exploitation costs due to unused of disc brakes for their entire lifetime. On the other hand, there may be situations where disc brakes reach maximum wear before a periodic revision is scheduled, because of some exploitation conditions that require a more intense use of the braking system of the vehicle.

Due to the complexity of the phenomena that occur in the exploitation of the braking system, phenomena that are influenced by different factors [7] and are characterized by some uncertainties, an analytical mathematical modeling of braking process it is difficult to be developed [1]. During the braking process, vibrations are produced within the contact area between the discs and brake pads and their spectrum depends on factors such as the degree of wear and temperature. Studies on vibrations in the car braking system are presented in ([13], [15], [16]), while investigations about the temperature of disc brakes are shown in ([3], [8], [23]).

The employement of fuzzy logic in automotive engineering has been presented in different studies. Von Altrock [20] emphasized the use of fuzzy logic for the ABS systems, control of the engine and control of automatic transmission. The control of the braking system, the suspension control and the vehicle dynamics control are presented among the relevant research areas of fuzzy control in automotive field ([9]). The use of fuzzy logic for the ABS systems has been shown

in other studies ([5], [11], [21], [22]). Fuzzy logic was also used to improve the handling and stability of vehicles ([4], [18]), to control the speed of an automotive engine ([19]) or to control a suspension system of automotive ([6]).

While fuzzy logic in automotive field has been used lately, its employement for the estimation of technical state of the disc brakesis is still a challenging approach and this area of research is less developed. Within this framework, the fuzzy logic ([12], [14], [17]) has been proposed for the estimation of technical state of the disc brakes, so they can be replaced before the maximum limit of wear is reached and without depending on the planned periodic revisions.

The research begins with the description of the experimental stand developed for the estimation of the technical condition of disc brakes. A fuzzy logic decision approach for the estimation of technical state of the disc brakes is shown in the next section. The implementation of the fuzzy logic decision approach for the estimation of technical state of the disc brakes with the Fuzzy Logic Toolbox ^{TM} of the Matlab ® software is also depicted. At the end, conclusions and recommendations for future research are presented.

2 Experimental method

For the estimation of the technical condition of disc brakes, the stand depicted Fig. 1 has been developed. The stand is based on a powertrain that consists of the following main elements:
a) four cylinders diesel engine with the following characteristics: maximum power 43.5 kW at 4600 rot/min, maximum torque 110 Nm at 2000 rot/min;
b) gearbox 5-speed manual transmission;
c) MacPherson front suspension;
d) hydraulic brake system composed of:
- a master brake cylinder, operated by a screw to ensure precise movement of the piston within its interior;
-a brake pipe and a brake caliper;
-a full disc brake with the diameter of 247 mm;
-brake pads.

The FLIR SC 640 infrared camera was used to monitor the temperature of braking disc. The vibration monitoring was performed with a system composed of PC-TopMessage device (made by Delphin Technology AG)-accelerometer sensor. The fluid pressure in the hydraulic circuit was monitored through a manometer with glycerin (scale 0 ... 6 MPa). The speed rotation of the disc brake (expressed in km/h) was monitored by a speedometer.

The contact force between the brake pads and disc was kept constant by maintaining the pressure of the brake pipe fluid at 1 MPa. The disc rotation speed was also maintained constant at a value corresponding to vehicle speeds of 30 km/h. The first speed of the gearbox was used to provide an engine speed in the stable operation area, without trepidation that could affect the vibrations measurement. Due to the dilatation of the material of the disk brake as a result of increasing its temperature during the braking process, we proceeded to fine tune of the position of the piston in the brake master cylinder using a screw-nut mechanism to maintain constant the pressure of the brake fluid.

The experiments were conducted using two disc brakes, corresponding to a new disc (thickness = 10 mm) and to a disc with the thickness at the lower limit prescribed by the manufacturer (thickness = 8.5 mm).

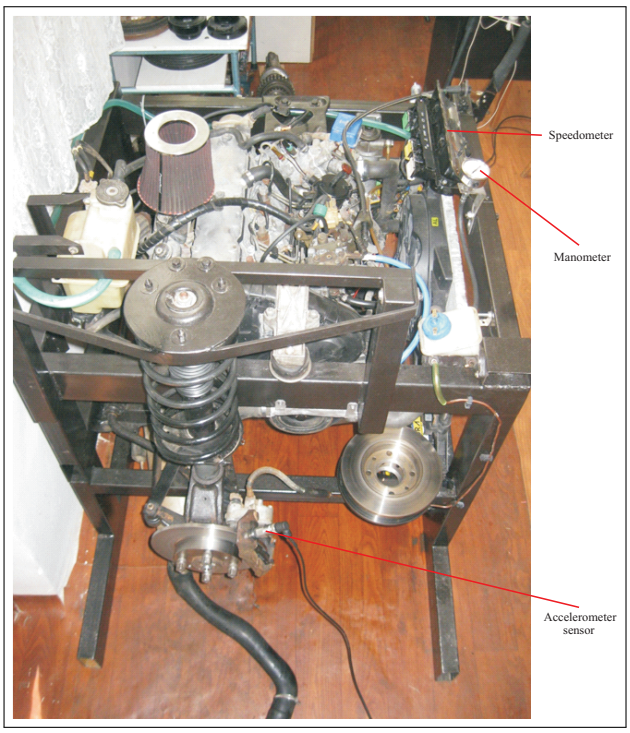


Figure 1: The experimental stand for the estimation of the technical condition of disc brakes

3 Results

A lowest value of the vibration amplitude equal to 1.46 mm/s rms was measured for a new disc brake (thickness=10 mm) at a temperature of 50.9 ^{0}C. A highest value of the vibration amplitude equal to 2.4 mm/s rms was measured for a disc brake with the thickness of 8.5 mm at a temperature of 154.2 ^{0}C.

The fuzzy logic decision approach described in [2, p.22-23] was employed for the estimation of technical state of the disc brakes, as follows:
1) The T=temperature [^{0}C] and Avib=Vibration's amplitude [mm/s] rms were used as the input linguistic variables:

$$LV_I = \{T, A_{vib}\} \tag{1}$$

Considering the measured value of the T and Avib, their domain values have been defined as follows:

$$\begin{aligned} T : D_T &= [50.9, 154.2] \\ A_{vib} : D_{A_{vib}} &= [1.46, 2.4] \end{aligned} \tag{2}$$

2) For the T and Avib variables the linguistic terms have been established as:

$$T : LT^T = \left\{ \begin{matrix} Tverysmall \\ Tsmall \\ Tmedium \\ Tbig \\ Tverybig \end{matrix} \right\} \tag{3}$$

and

$$Avib : LT^{Avib} = \begin{Bmatrix} AVverysmall \\ AVsmall \\ AVmedium \\ AVbig \\ AVverybig \end{Bmatrix} \tag{4}$$

Among the membership functions, the triangular functions are extensively employed and they have been proposed as the membership functions for both T and Avib variables. According with the definition of [10, p.25], their expressions are:

$$mf_k^T(x, a_{T_k}, b_{T_k}, c_{T_k}) = \begin{cases} 0, & x \le a_{T_k} \\ \frac{x - a_{T_k}}{b_{T_k} - a_{T_k}}, & a_{T_k} \le x \le b_{T_k} \\ \frac{c_{T_k} - x}{c_{T_k} - b_{T_k}}, & b_{T_k} \le x \le c_{T_k} \\ 0, & c_{T_k} \le x \end{cases}, k = 1...5 \tag{5}$$

and

$$mf_j^{Avib}(x, a_{Avib_j}, b_{Avib_j}, c_{Avib_j}) = \begin{cases} 0, & x \le a_{Avib_j} \\ \frac{x - a_{Avib_j}}{b_{Avib_j} - a_{Avib_j}}, & a_{Avib_j} \le x \le b_{Avib_j} \\ \frac{c_{Avib_j} - x}{c_{Avib_j} - b_{Avib_j}}, & b_{Avib_j} \le x \le c_{Avib_j} \\ 0, & c_{Avib_j} \le x \end{cases}, j = 1...5 \tag{6}$$

where:

- $a_{T_k} < b_{T_k} < c_{T_k}$ are the parameters of each triangular membership function mf_k^T in the expression (5), k=1...5;

- $a_{Avib_j} < b_{Avib_j} < c_{Avib_j}$ represent the parameters of each triangular membership function mf_j^{Avib} in the expression (6), j=1...5.

3)The TBD=the thickness of the brake disc [mm] was employed as the output variable:

$$LV_O = \{TBD\} \tag{7}$$

For the TBD variable, the domain value has been defined as:

$$TBD : D_{TBD} = [8.5, 10] \tag{8}$$

4) For the TBD variable the linguistic terms have been established as:

$$TBD : LT^{TBD} = \begin{Bmatrix} TBDverysmall \\ TBDsmall \\ TBDmedium \\ TBDbig \\ TBDverybig \end{Bmatrix} \tag{9}$$

The triangular function have also been proposed as the membership function for TBD variable and its expression is [10, p.25]:

$$mf_r^{TBD}(x, a_{TBD_r}, b_{TBD_r}, c_{TBD_r}) = \begin{cases} 0, & x \le a_{TBD_r} \\ \frac{x - a_{TBD_r}}{b_{TBD_r} - a_{TBD_r}}, & a_{TBD_r} \le x \le b_{TBD_r} \\ \frac{c_{TBD_r} - x}{c_{TBD_r} - b_{TBD_r}}, & b_{TBD_r} \le x \le c_{TBD_r} \\ 0, & c_{TBD_r} \le x \end{cases}, r = 1...5 \quad (10)$$

where $a_{TBD_r} < b_{TBD_r} < c_{TBD_r}$ represent the parameters of each triangular membership function mf_r^{TBD} in the expression (10), r=1...5.
5) The fuzzy rules base is:

$$\begin{aligned} Rule_1 &: \{IF\ (AVverysmall)\ and\ (Tverysmall)\ then\ (TBDverybig)\} \\ Rule2 &: \{IF\ (AVverysmall)\ and\ (Tsmall)\ then\ (TBDverybig)\} \\ Rule_3 &: \{IF\ (AVsmall)\ and\ (Tverysmall)\ then\ (TBDverybig)\} \\ &... \\ Rule_9 &: \{IF\ (AVmediu)\ and\ (Tmedium)\ then\ (TBDmedium)\} \\ &... \\ Rule_{23} &: \{IF\ (AVverybig)\ and\ (Tbig)\ then\ (TBDverysmall)\} \\ Rule_{24} &: \{IF\ (AVbig)\ and\ (Tverybig)\ then\ (TBDverysmall)\} \\ Rule_{25} &: \{IF\ (AVverybig)\ and\ (Tverybig)\ then\ (TBDverysmall)\} \end{aligned} \quad (11)$$

6) The centroid method [17, p. 98] has been used to obtain the the defuzzified value TBD of the thickness of the disc brake.

The fuzzy logic decision approach for the estimation of technical state of the disc brakes has been implemented with the Fuzzy Logic Toolbox ^{TM} of the Matlab ® software (Fig. 2). The inference rules are shown in figure 3. Figure 4 depicts the dependence of the TBD=f(T, Avib). Considering T=116 ^{0}C and Avib=2.16 mm/s rms we obtain TBD=8.9 mm, so that thickness of the disc brake is within prescribed limits [8.5,10].

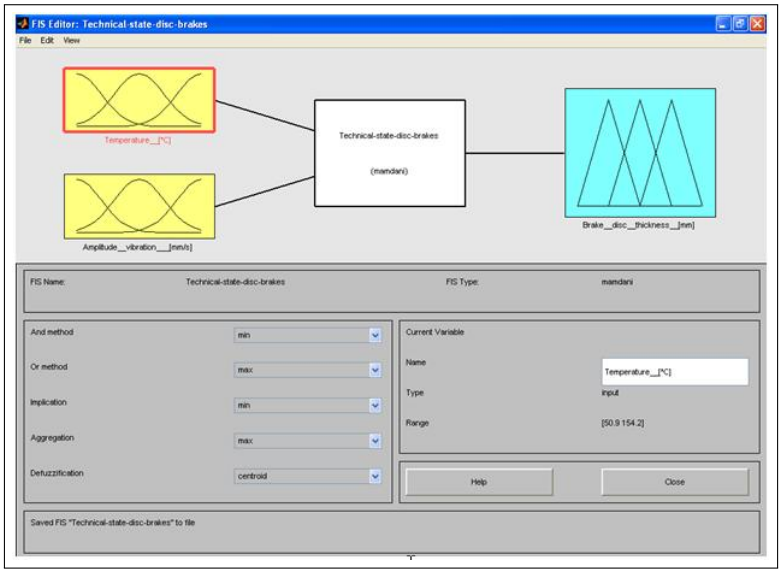


Figure 2: The fuzzy decision system Technical-state-disc-brakes.fis

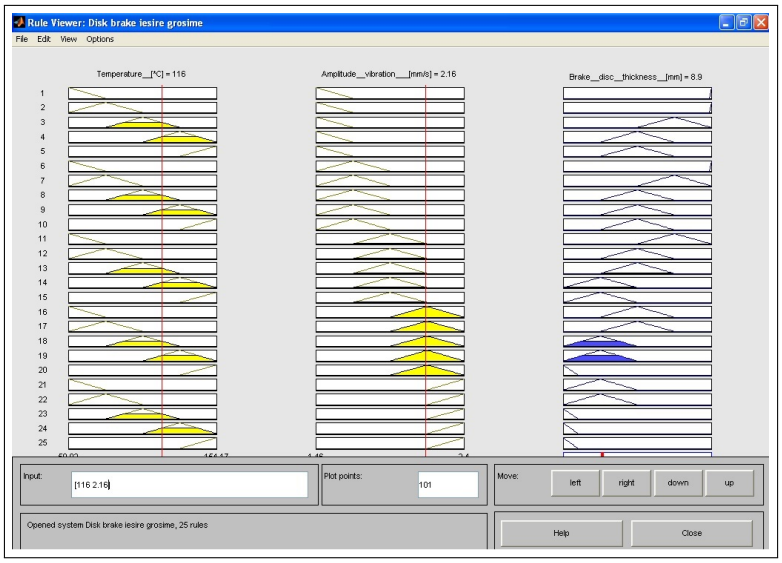


Figure 3: The inference rules of the fuzzy decision system Technical-state-disc-brakes.fis

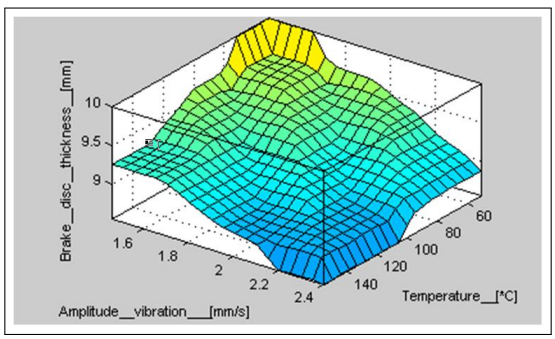


Figure 4: The dependence TBD=f(T, Avib)

4 Conclusions

The phenomena that occur in the exploitation of the braking system are very complex and they are characterized by some uncertainties. Therefore, a fuzzy logic approach has been employed in this research for the estimation of technical state of the disc brakes, expressed through their thickness variation.

A stand has been developed to establish the domain values of the temperature and vibration amplitude of the disc brakes, which were used as the input linguistic variables. The thickness of the disc brakes was employed as the output linguistic variable. The Fuzzy Logic Toolbox ^{TM} of the Matlab ® software was used to develop the fuzzy decision system, which can be applied to establish if the thickness of the disc brakes is within prescribed limits by the manufacturer.

In our research both the contact force between the brake pads and disc, and the disc rotation speed were maintained constant. Future studies are needed to investigate the development of a fuzzy decision system when these two characteristics are also variable. The use of fuzzy logic approach in joining with other artificial intelligence methods, including neural networks may also represent important area for future research.

Bibliography

[1] Aleksendric, D.; Barton, D. C. (2009); Neural network prediction of disc brake performance, *Tribology International*, 42(7): 1074-1080.

[2] Baban, M.; Baban, C.F.; Blaga, F.S. (2010); Maintenance planning of cold plastic deformation tools using fuzzy logic, *Eksploatacja i Niezawodnosc- Maintenance and Reliability*, 3: 21-26.

[3] Belhocine, A.; Bouchetara, M. (2012); Thermomechanical modelling of dry contacts in automotive disc brake, *International Journal of Thermal Sciences*, 60: 161-170.

[4] Boada, M.J.L.; Boada, B.L.; Munoz, A.; Diaz, V. (2006); Integrated control of front-wheel steering and front braking forces on the basis of fuzzy logic, *Proceedings of the Institution of Mechanical Engineers 220.D3 (Mar 2006)*, 253-267.

[5] Cabrera, J.A.; Ortiz, A.; Castillo, J.J.; Simon, A. (2005); A Fuzzy Logic Control for Antilock Braking System Integrated in the IMMa Tire Test Bench, *IEEE Transactions on Vehicular Technology*, 54(6):1937- 1949.

[6] Cherry, A.S.; Jones, R.P.(1995); Fuzzy logic control of an automotive suspension system, *IEE Proceedings: Control Theory and Applications*, 142(2):149-160.

[7] Cirovic, V.; Aleksendric, D. (2011); Dynamic modelling of disc brake contact phenomena, *FME Transactions*, 39(4):177-183.

[8] Hwang, P.; Wu, X. (2010); Investigation of temperature and thermal stress in ventilated disc brake based on 3D thermo-mechanical coupling model, *Journal of Mechanical Science and Technology*, 24:81-84.

[9] Ivanov, V. (2010); Fuzzy Methods in Ground Vehicle Engineering: State-of-the-Art and Advanced Applications, *Proceedings of the 8th International Conference on Structural Dynamics, EURODYN 2011 Leuven, Belgium, 4-6 July 2011, G. De Roeck, G. Degrande, G. Lombaert, G. Muller (eds.)*, 3008-3015.

[10] Jang, J.-S.R.; Sun, C. T.; Mizutani, E. (1997); *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, Upper Saddle River, NJ: Prentice-Hall.

[11] Mauer, G.F. (1995); A fuzzy logic controller for an ABS braking system,*IEEE Transactions on Fuzzy Systems*, 3(4):381-388.

[12] Ross, T. (2004); *Fuzzy Logic with Engineering Applications*, John Wiley & Sons Ltd., second edition.

[13] Sawczuk, W. (2011); Application of vibroacoustic signal to diagnose disck braking system, *Journal of KONES Powertrain and Transport*, 18 (1): 525-534.

[14] Seising, R. (2011); From Electrical Engineering and Computer Science to Fuzzy Languages and the Linguistic Approach of Meaning: The non-technical Episode: 1950-1975, *International Journal of Computers Communications & Control*, 6(3): 530-561.

[15] Sergienko, V. P.; Bukharov, S. N. (2009); Vibration and Noise in Brake Systems of Vehicles. Part 2: Theoretical Investigation Techniques, *Journal of Friction and Wear*, 30(3): 216-226.

[16] Sergienko, V. P.; Bukharov, S. N.; Kupreev, A. V.(2008); Noise and Vibration in Brake Systems of Vehicles. Part 1: Experimental Procedures, *Journal of Friction and Wear*, 29(3):234-241.

[17] Sivanandam, S.N.; Sumathi, S.; Deepa, S.N. (2006); *Introduction to Fuzzy Logic using MATLAB*, Springer.

[18] Song, J.(2012); Integrated control of brake pressure and rear-wheel steering to improve lateral stability with fuzzy logic, *International Journal of Automotive Technology*, 13(4): 563-570.

[19] Vachtsevanos, G.J.; Farinwata, S.S.; Pirovolou, D.K. (1993); Fuzzy logic control of an automotive engine, *IEEE Control Systems*, 11(3): 62-68.

[20] von Altrock, C.(1997); Fuzzy Logic in Automotive Engineering, *Circuit Cellar INK, The Computer Applications Journal*, 88: 1-9.

[21] Will, A.B., Zak, S.H. (2000); Antilock brake system modelling and fuzzy control, *International Journal of Vehicle Design*, 4(1): 1-18.

[22] Yazicioglu, Y.; Unlusoy, Y.S. (2008); A fuzzy logic controlled Anti-lock Braking System (ABS) for improved braking performance and directional stability, *International Journal of Vehicle Design*, 48(3-4): 299-315.

[23] Yevtushenko, A.; Kuciej, M. (2010); Temperature and thermal stresses in a pad/disc during braking, *Applied Thermal Engineering*, 30: 354-359.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 9(5):539-554, October, 2014.

ANN Method for Control of Robots to Avoid Obstacles

E. Ciupan, F. Lungu, C. Ciupan

Emilia Ciupan, Florin Lungu, Cornel Ciupan*
Technical University of Cluj-Napoca
Romania, 400641 Cluj-Napoca, Bd. Muncii, 103-105
emilia.ciupan@mis.utcluj.ro, florin.lungu@mis.utcluj.ro
*Corresponding author: cornel.ciupan@muri.utcluj.ro

Abstract: The avoidance of obstacles placed in the workspace of the robot is a problem which makes controlling them more difficult. The known avoidance methods used for the robots control are based on bypass trajectory programming or on using the sensors that detect the position of the obstacle. This paper describes a method of training industrial robots in order for them to avoid certain obstacles in the workspace. The method is based on the modelling of the robot's kinematics by means of an artificial neural network and by including the neural model in the robot's controller. The neural model simulates the robot's inverse kinematics, and provides the joint coordinates, as referential values for the controller. The novelty of the method consists in the deliberately erroneous training of the network, so that, when programming a direct trajectory in the workspace, the robot avoids a known obstacle.
Keywords: Artificial Neural Network (ANN), control, robot, obstacle avoidance.

1 Introduction

Proportional-integrative-derivative PID controller is widely used for the control of robots, because it is model-free, and its parameters can be adjusted easily and separately. An integrator in a PID controller reduces the bandwidth of the closed-loop system. In order to remove steady-state error caused by uncertainties and noise, the integrator factor has to be increased, having the effect of reducing the performance of transient regime [19].

The application of neural networks to robots control is well known [10], [11] and an alternative to the adaptive control is represented by the neural controllers [21].

Lewis et al. [10] demonstrate that neural networks do indeed fulfil the promise of providing model-free learning controllers for a class of nonlinear systems. Neural network control offers two specific advantages over adaptive control:

- neural network controller works for any rigid robot arm without computing a regression matrix or performing any preliminary analysis
- neural networks provide a basis set for any smooth function, while the linear in the parameters equation provides a basis set only for linear systems.

There are several approaches to combine PID control with the intelligent control, such as the neural control. The first way is to form neural networks into PID structure [5], [6], [10], [17]. By proper updating laws, the parameters of PID controllers are changed so that the closed-loop systems are stable. The second method used intelligent techniques to tune the parameters of PID controllers, such as fuzzy tuning [11], neural tuning [7], [18], and expert tuning [8].

All known approaches require the set point (the reference values) determination that consists in drive joints coordinates which are obtained from the inverse kinematic analysis.

In the inverse kinematic analysis [14], [16], the coordinates and the effector's orientation (X,Y,Z,ψ,θ,φ) are considered to be known, and the coordinates of the joints (represented by q_i, i=1, ..., m, where m is the number of kinematic axes, equal to the number of the degrees of

freedom) are to be determined. Although an apparently easy task, determining the coordinates of the joints becomes more complicated when robots with complex kinematic structure, such as the parallel robots, are at stake [20].

Given its advantages, neural computing is often used to solve the problem of inverse kinematics. The training of the neural network and the getting of the neural model implies solving two important problems [7]:

a) getting the training data, especially when the mathematical model is not known, and measurements on the physical model are necessary

b) performing the training process and obtaining an acceptable error, in the case of a large amount of training data.

The proper control of the robot is carried out by the robot's control equipment, by means of generating a control input for each joint, so that it achieve coordinate q_i resulted from the inverse kinematics, and the effector pass through the points that belong to the trajectory. Therefore an important problem is to determine the coordinates of joints.

Figure 1 shows a neural controller for the positioning of the effector, which uses a neural model NM for the generation of the coordinates of the joints q_i, PID controllers and feedback loops. The method implies the completion of the following stages: providing the coordinates of the some points that define the robot's trajectory, generating some additional points on this trajectory and determining the coordinates of the joints by using an NM neural model implemented onto the robot's control equipment, transmitting the coordinates of the joints to the controllers of the PID_i axes, which generate the actuating quantity e_i corresponding to the M_i motors of the robot.

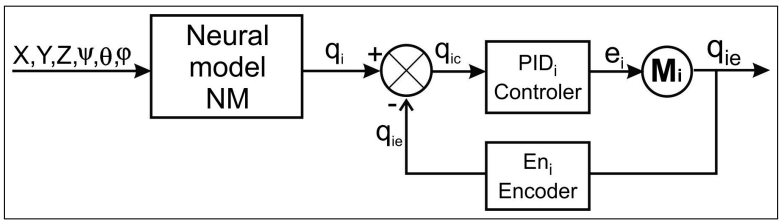


Figure 1: Neural network controller

In order to obtain and implement the NM neural model, it is necessary to complete the following steps: creating a neural network, creating a set of training examples, training the neural network, which results in the creation of the neural model, testing and validating the neural model, and using the neural model by implementing it in the control equipment. The set of training examples consists of pairs of input-output data which are determined by the choice of a point cloud in the robot's workspace. The input signals are considered to be the positioning coordinates of the points in the workspace, while the output signals, the coordinates of the joints associated with these positions.

The disadvantage of this method is that when programming a straight trajectory between two points the robot cannot avoid an obstacle, even though the volume covered by the obstacle has been excluded from the set of training examples. At the same time, another disadvantage consists in the large number of training examples needed to cover the entire workspace of the robot.

In the majority of the handling applications, the robot's task is to complete linear movement between points that belong to the workspace, points where it has to retrieve or deliver objects, or where the robot perform operations. This general method can also be used to obtain a neural

model in the case of the effector's move between two points, so that it can avoid an obstacle. In this case, the set of training examples is constituted by the association of coordinates of the points on the deviation trajectory with the coordinates of the joints corresponding to them. It is necessary for the trajectory to be described in this case.

A way of describing the trajectory is by its mathematical expression, which has the disadvantage of having to determine it. The next step consists in the determination of the coordinates of the joints for a set of points that belong to the bypass trajectory. For complex structure robots, such as the parallel robots with 6 degrees of freedom, the expression of the joints determination is complicated. The general form of the expression is $q_i=f_i(X,Y,Z,\psi,\theta,\varphi)$, i=1, ...,6. In this case, the calculation is complex and it involves a large number of mathematical operations. This is a disadvantage, especially in the case of robots that operate at high speeds, given that calculations are made in real time.

Also, in the patent literature [9], [12], [13], there are many methods of robot control using neural networks. Patent CN102346489 discloses a pulse neural network method of robot object tracking control. The collision avoidance is done by processing a set of information from the sensors. There is no information regarding a method to avoid a static obstacle in the robot's workspace to be achieved exclusively by neural network training, without the use of visual sensors to identify the obstacle [15].

Some of the authors have been previously involved in research concerning the use of neural networks for economic applications, or robot control. The program presented in [1], [2] was designed with the purpose of using neural computing in the modelling and the simulation of processes or activities. It is suitable for the study of any activity for which a three-layer perceptron neural network may serve as a model.

Also, [3] shows a neural model for the kinematical analysis of six parallel robot. For reasons related to simplification, there has been considered the move of the effector in a cube with a side of 10 cm, without taking into account the variation of the position angles. For the training of the network, there have been generated 130 training examples, and then the neural model for the move of the robot on different trajectories has been validated. In the application of the neural model, there has been noticed that the training of the neural model in a larger working space, specific to a robot, is difficult, especially when the robot has to avoid an obstacle. That is why the authors have aimed to develop a more effective method of control for the cases in which the robot has to avoid an obstacle.

In [4] there has been presented the main principle of a method of obstacle avoidance, by means of an erroneous instruction of the network. The experiments have been completed in a smaller part of the robots workspace (a cube with the side 10 cm). Further research illustrated the difficulty of obtaining an effective neural model that can lead the effector with a precision that is appropriate to the application, and avoid the obstacle. There have been circumstances in which the neural model did not offer the coordinates that lead the effector beside the obstacle [3]. The validation of the models presented [3], [4] has been made based on sets of data that have not been used as training data; there has been completed no cross-validation.

In the current paper the authors have developed the method in terms of the generation of the set of the training examples. In this sense, there have been stated clear rules that establish the training examples. In a first stage, there has been completed a neural model tested through a 4-fold cross-validation technique. The case studies have been carried out in a workspace that had the shape of a cube with a side of 600 mm, corresponding to the majority of the applications, by using an obstacle with a cube shape with a side of 200 mm. In order to avoid the collision, there have been studied three types of envelopes.

There are several types of neural networks that can be used in modelling a system. They can be classified based on criteria such as the structure or the instruction types (networks of

perceptron type, radial basis function networks, Kohonen self-organization networks, Hopfield networks, fuzzy neural networks, networks with supervised or non-supervised training and so on). In all the modelling activities that are referred to in this paper there have been used neural networks of the type of a three-layer perceptron, in which the initial layer has 6 neurons corresponding to the position (X,Y,Z,ψ,θ,φ) of the effector. In the case of some of these models, the final layer has 6 neurons corresponding to the coordinates of the joints q_i, i=1, ...,6, or only one neuron, corresponding to the coordinate of a single joint (q_3 for instance). The number of the neurons of the hidden layer has been determined by trial, throughout several training sessions, based on the criteria of the minimization of the mean square error. The chosen activation functions have been log-sigmoid for the neurons of the hidden layer, and the function purelin for the neurons on the output layer. The training of the networks has been completed using the Levenberg-Marquardt method. Unlike the descent gradient method, the process of training through the Levenberg-Marquardt method could converge quickly when close to the solution. As it is a method based on Hessian, there is no risk that in such a circumstance the solution be lost, as it can occur in the training with the descent gradient method, when a higher rate of learning is used. In the modelling activity, there has been used the Matlab application.

2 Method to avoid obstacle

2.1 Method presentation

The problem that this paper solves consists in the elaboration of an industrial robot training method based on neural network modelling and training, so that, when programming a straight trajectory between two points through which the robot's effector has to pass, the robot avoid an obstacle that it encounters.

The method of training robots to avoid obstacles is based on the modelling, training and use of three-layer perceptron type neural networks, having k neurons in the input layer, which corresponds to the number of degrees of freedom, m neurons in the output layer, which corresponds to the number of kinematic axes, and a number n, consisting of 15 to 50 neurons, which corresponds to the hidden layer. Figure 2 shows the scheme of a neural network of this type that has 6 neurons in the input layer, corresponding to the 6 degrees of freedom, 6 neurons in the output layer, corresponding to the 6 kinematic axes and an unspecified number n, in the hidden layer.

The training data are determined from the mathematical model or by experimenting on the physical model of the robot, by choosing a point cloud contained in a work plane included in the robot's workspace, plane in which the robot has to operate at least one move between two given points. The avoidance of the obstacle in the robot's trajectory is achieved by the training of the network, with input data corresponding to the coordinates of some points on the robot's direct trajectory between two points, and output data (the joints coordinates) corresponding to the bypass trajectory.

The set of training examples will have as input signals the coordinates of the cloud of points, while as output ones, the coordinates of the joints, calculated according to the rule R below:

R1) for a point in the work plane outside the obstacle or its envelope, it is established as pair the coordinates of the correct joints (in accordance with the mathematical model which describes the robot's kinematics, or in accordance with the experimental measurements)

R2) for a point in the work plane that belongs to the obstacle or its envelope, it is established as pair the coordinates of the joints that belong to a different point, the latter being situated on the surface of the obstacle, or on its envelope, depending on the case.

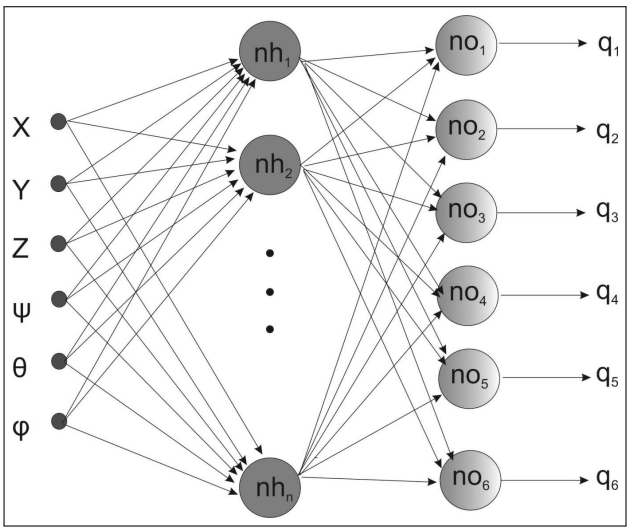


Figure 2: Neural network diagram

The number of neurons in the hidden layer is chosen by means of trials, a practice which is used in neural computing.

The novelty of this method consists in the way the training set is built. This leads to a deliberately erroneous training, so that in the recall phase it is not necessary to know the bypass trajectory.

The set of training examples is constituted by input-output data pairs, in which the input signals correspond to some points on the robot's direct Td trajectory. The output signals, in the training phase, are the coordinates of the joints, but in "deliberately erroneous" way, they are not the coordinates associated to the direct trajectory that crosses the obstacle, but to the output signals corresponding to the points situated outside the obstacle, on a Ta avoidance trajectory. Thus, in the recall phase, the model is going to behave erroneously. This means that for the points on a direct trajectory (a line, in most of the cases) which does not bypass the obstacle, transmitted as input data, the neural model will generate, as output, the coordinates of the joints that will lead the robot beside the obstacle.

In order to achieve the neural model NM, one has to complete the following steps:

- create a neural network that has, in its input layer, a number of neurons equal to the number of the robot's degrees of freedom, and in its output layer, a number of neurons equal to the number of joints q_i

- create a set of training examples formed by pairs of effector coordinates, which belong to the robot's work plane, and corresponding coordinates of the joints q_i, determined according to rule R described above

- train a neural network with the sets of training data, the result of the training process being called "neural model"

- test the neural model achieved previously and validate it, in case acceptable errors are obtained

- use the neural model, which means that the neural model receives exclusively input data which consists of robot effector positions, and generates the coordinates of the joints.

2.2 Cross validation of the method

Let us consider the case of a serial robot with six degrees of freedom (Figure 3), given by three positioning movements (X, Y and Z), and other three effector orientation movements (ψ, θ, φ). Based on the robot's kinematic scheme (Figure 3a), one can describe the connecting relations

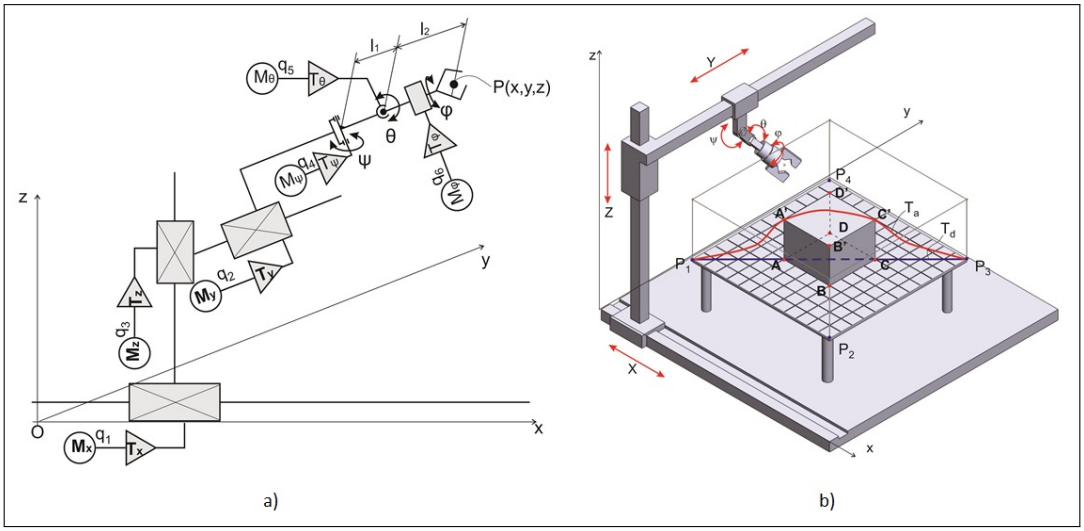


Figure 3: a. Kinematic diagram; b. The robot architecture

between the joints coordinates q_i, as well as the effector's position (X, Y, Z, ψ, θ, φ). Thus, the mathematical model for the inverse kinematics is obtained by solving the system of equations (1)-(6):

$$X = X_0 + q_1 \cdot i_{Tx} + l_2 sin(\theta) \tag{1}$$

$$Y = Y_0 + q_2 \cdot i_{Ty} + (l_1 + l_2 cos(\theta)) sin\psi \tag{2}$$

$$Z = Z_0 + q_3 \cdot i_{Tz} - (l_1 + l_2 cos(\theta)) cos\psi \tag{3}$$

$$\psi = \psi_0 + q_4 \cdot i_{T\psi} \tag{4}$$

$$\theta = \theta_0 + q_5 \cdot i_{T\theta} \tag{5}$$

$$\varphi = \varphi_0 + q_6 \cdot i_{T\varphi} \tag{6}$$

where i_{Tx}, i_{Ty}, i_{Tz}, $i_{T\psi}$, $i_{T\theta}$, $i_{T\varphi}$ represent the transfer functions of the transforming mechanisms which generate the given movements, and X_0, Y_0, Z_0, ψ_0, θ_0, φ_0 represent the initial values obtained for q_i=0, i=1,...,6.

The architecture of the robot in Figure 3a is shown in Figure 3b. A portion of the robot's workspace (Figure 3b), which has the shape of a parallelepiped with base $P_1P_2P_3P_4$ is being considered. In the workspace there is an obstacle ABCD$A'B'C'D'$ which has to be avoided during the operation. It is assumed that the robot's effector has to move between points P_1 P_3, but on a trajectory that should avoid the obstacle ABCD$A'B'C'D'$.

The plane of points P_1, P_2, P_3, P_4 is called work plane (represented as Wp), namely a plane in which the robot has to complete a number of operations (retrieve or place objects, feed equipment etc.).

We are looking for a simple and comfortable method of the robot control, which would provide as input data the coordinates of some points on the direct trajectory Td and make the robot move

on the trajectory Ta which avoids the obstacle. This is achieved by a "deliberately erroneous" training of the neural network which models the robot's behaviour.

The set of training data, under the form of input-output matrix pairs, is obtained by the association of the effector's coordinates (X, Y, Z, ψ, θ, φ) with the joints coordinates (q_i, i=1, ..., 6) resulted by means of mathematical model or by measuring on the physical model. The generation of the training data is made according to table of Figure 4.

No.	Points of training	Wished input						Wished output	Origin q_i
1.	P_1	X_1	Y_1	Z_1	ψ_1	θ_1	φ_1	$q_{i,1}$	P_1
2.	P_2	X_2	Y_2	Z_2	ψ_2	θ_2	φ_2	$q_{i,2}$	P_2
3.	P_3	X_3	Y_3	Z_3	ψ_3	θ_3	φ_3	$q_{i,3}$	P_3
4.	P_4	X_4	Y_4	Z_4	ψ_4	θ_4	φ_4	$q_{i,4}$	P_4
5.	A	X_A	Y_A	Z_A	ψ_A	θ_A	φ_A	$q_{i,A'}$	A'
6.	B	X_B	Y_B	Z_B	ψ_B	θ_B	φ_B	$q_{i,B'}$	B'
7.	C	X_C	Y_C	Z_C	ψ_C	θ_C	φ_C	$q_{i,C'}$	C'
8.	D	X_D	Y_D	Z_D	ψ_D	θ_D	φ_D	$q_{i,D'}$	D'
9.	O_1	X_{O1}	Y_{O1}	Z_{O1}	ψ_{O1}	θ_{O1}	φ_{O1}	$q_{i,O'1}$	O'_1

Figure 4: The way for obtaining the training data

In order to avoid the obstacle, the robot has to move on a deviating trajectory Ta. For the robot to move on the deviating trajectory Ta, its control equipment has to receive information regarding the shape of the trajectory as some coordinates of some points on the trajectory. A possible description of the trajectory is given by its mathematical expression, which has the disadvantage of having to determine it. The next step is to determine the coordinates of the joints for a set of points which belong to the deviating trajectory Ta.

According to the approach of this paper the avoiding trajectory is approximated by a set of few points, without knowing the mathematical expression of the trajectory, and the determination of the joints is done on basis of a neural model.

In order to complete the data in table of Figure 4, in the case of the robot in Figure 3b, there have been chosen the points P_1(200,200,200), P_2(800,200,200), P_3(800,800,200) and P_4(200,800,200). The obstacle is considered to be a parallelepiped defined by the points A(400,400,200), B(600,400,200), C(600,600,200) and D(400,600,200), situated in the work plane and the points A'(400,400,400), B'(600,400,400), C'(600,600,400) and D'(400,600,400), situated in a plane parallel to the work plane, 200 mm away.

For the validation of the method, there has been considered the move of the effector on the diagonal P_1P_3 (figure 3b). In order to establish the set of the training and testing examples for the move on the segments P_1A and CP_3, rule R1 is applied. As for the movement on the segment AC, there is applied rule R2. The set of the training data is shown in table of Figure 5.

There has been applied a 4-fold cross-validation technique of the method. Thus, the set of the training examples in table of Figure 5 has been divided into four subsets. The inclusion rule regarding the training examples in the four subsets is described by means of the indexes of the lines in table of Figure 5, grouped in the subsets M_k, k=1, ...,4, according to the algorithm below:

$$M_k = \{j_{k,n} \mid j_{k,n} = k + 4n, k \in [1, 4], n \in [0, 75]\} \tag{7}$$

For the validation of the method, out of the four subsets, there has been successively retained a subset for validation, while the other three subsets have been used, merged, for training. Thus, there have been completed four rounds of training and validation of the method, using the neural networks. Within each round, there have been developed several neural models of the three-layer perceptron type, having the architecture 6-m-6, where m represents the number of

Training points	Effector coordinates							Joint coordinates ($q_{i,j}$, i=1,...,6; j=1,...,301)					
j	X_j	Y_j	Z_j	Z'_j	ψ_j	θ_j	φ_j	$q_{1,j}$	$q_{2,j}$	$q_{3,j}$	$q_{4,j}$	$q_{5,j}$	$q_{6,j}$
0	1	2	3	4	5	6	7	8	9	10	11	12	13
1	200	200	200	200	0	0	0	20	20	20	20	20	20
2	202	202	200	200	0	0	0	20.2	20.2	20	20	20	20
3	204	204	200	200	0	0	0	20.4	20.4	20	20	20	20
4	206	206	200	200	0	0	0	20.6	20.6	20	20	20	20
...													
100	398	398	200	200	0	0	0	39.8	39.8	20	20	20	20
101	400	400	200	400	0	0	0	40	40	40	20	20	20
102	402	402	200	400	0	0	0	40.2	40.2	40	20	20	20
103	404	404	200	400	0	0	0	40.4	40.4	40	20	20	20
...					0	0	0	0	0	0	20	20	20
199	596	596	200	400	0	0	0	59.6	59.6	40	20	20	20
200	598	598	200	400	0	0	0	59.8	59.8	40	20	20	20
201	600	600	200	400	0	0	0	60	60	40	20	20	20
202	602	602	200	200	0	0	0	60.2	60.2	20	20	20	20
203	604	604	200	200	0	0	0	60.4	60.4	20	20	20	20
204	606	606	200	200	0	0	0	60.6	60.6	20	20	20	20
...													
299	796	796	200	200	0	0	0	79.6	79.6	20	20	20	20
300	798	798	200	200	0	0	0	79.8	79.8	20	20	20	20
301	800	800	200	200	0	0	0	80	80	20	20	20	20

Figure 5: The set of training examples for cross-validation

the neurons in the hidden layer. The activation functions chosen have been the log-sigmoid for the neuron for the hidden layers, and the purelin function, respectively, for the neuron on the output layer. The instruction has been completed by using the Levenberg-Marquardt method, in the case of the Matlab application. Following the criterion of the minimization of the mean square error throughout the repeated trainings, there have been retained models for which m=31. The training parameters have default values, namely maximum epochs (1000), performance goal (0), minimum gradient (10^{-05}), maximum validation checks (6), multiplication factor (0.001), multiplication factor decrease ratio (0.1), multiplication factor increase ratio (10), maximum value of multiplication factor (10^{10}).

Table presented in Figure 6 shows the most effective (minimum mean square error, marked as MSE) obtained when training the networks in the case of each of the four rounds.

Round	Training MSE
1.	$6.0xe^{-3}$
2.	$2.8xe^{-2}$
3.	$3.0xe^{-2}$
4.	$7.9xe^{-4}$

Figure 6: The best MSE

It has been noticed that for each of the four neural models obtained by means of the combination of three subsets M_k, there has been obtained, through testing on the fourth test subset, very good results for the coordinates X, Y, ψ, θ, φ. The results of the four simulations are briefly shown in Figure 7 and in table of Figure 8.

The analysis of the results obtained shows that there appear problems when simulating the coordinate Z at the intersection of the direct trajectory with the obstacle, close to the latter. Outside the area close to the points AA' and CC', all the four simulations grant good results for coordinate Z as well.

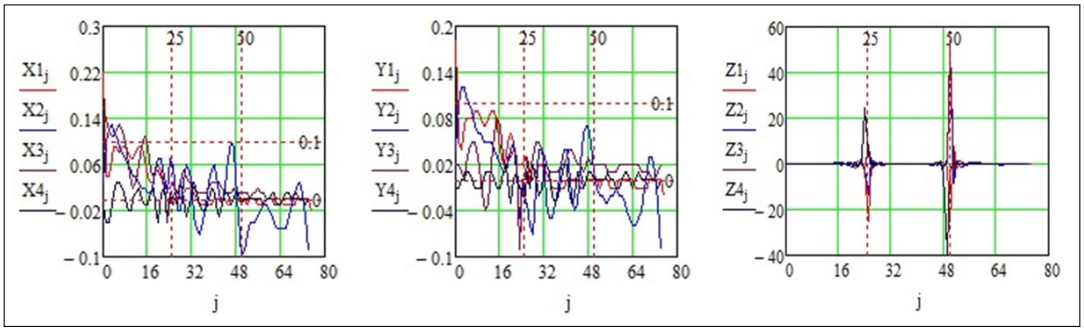


Figure 7: The errors of coordinates X, Y, Z

No. testing points	Cross validation											
j	M_1			M_2			M_3			M_4		
0	$\%X1_j$	$\%Y1_j$	$\%Z1_j$	$\%X2_j$	$\%X2_j$	$\%Z2_j$	$\%X3_j$	$\%Y3_j$	$\%Z3_j$	$\%X4_j$	$\%Y4_j$	$\%Z4_j$
1	0.22	0.18	-0.03	0.03	0.04	0.05	0.16	0.02	-0.08	-0.01	-0.01	0.02
2	0.04	0.05	0.05	0.10	0.10	0.07	0.14	0.02	-0.13	-0.04	-0.01	0.05
3	0.04	0.04	0.06	0.12	0.12	0.06	0.12	0.01	-0.17	-0.04	0.00	0.04
...												
21	0.06	0.06	0.01	0.07	0.04	-1.08	0.05	0.03	0.34	-0.04	-0.03	0.27
22	0.05	0.05	0.10	0.07	0.05	-0.72	0.04	0.04	-1.78	0.01	0.00	0.38
23	0.02	0.02	-0.47	0.04	0.01	1.58	0.03	0.03	-1.52	0.02	0.01	-1.08
24	0.02	0.02	0.83	0.01	-0.05	0.13	-0.04	-0.09	2.17	-0.03	-0.01	2.92
25	0.02	0.02	-2.49	0.07	0.00	-2.73	0.00	0.00	-3.40	0.01	-0.01	24.70
26	0.03	0.03	-24.97	0.06	-0.03	-13.56	0.00	0.00	8.70	-0.01	-0.02	3.85
27	0.02	0.02	1.26	0.03	-0.03	2.29	0.01	0.03	-3.00	0.00	0.01	-0.09
28	0.00	0.00	-0.64	-0.02	-0.06	-1.07	0.01	0.01	1.38	-0.01	-0.01	0.12
29	0.00	0.00	0.22	-0.02	-0.07	-0.15	0.03	0.03	0.12	0.03	0.02	-0.15
30	0.01	0.01	0.07	0.03	-0.01	0.40	0.05	0.05	-0.81	0.00	0.00	0.07
...												
46	-0.01	-0.01	-0.09	0.07	0.02	-0.57	0.01	0.02	0.21	0.01	0.01	-0.07
47	-0.01	-0.01	-0.33	0.10	0.06	0.39	0.01	0.01	0.33	-0.01	-0.01	-0.44
48	0.00	0.00	0.17	0.09	0.07	0.83	0.02	0.02	-0.41	0.00	0.00	1.05
49	0.00	0.00	0.78	0.04	0.07	-2.02	0.01	0.03	-0.41	0.01	0.01	-3.93
50	0.00	0.00	-1.49	-0.03	0.05	2.75	0.01	0.02	0.73	0.01	0.01	-39.31
51	0.00	0.00	-24.99	-0.10	0.01	47.89	0.01	0.02	51.64	-0.01	0.00	-3.01
52	0.00	0.00	7.55	-0.09	-0.01	-8.56	0.01	0.02	-3.20	0.00	0.00	0.64
53	-0.01	-0.01	-1.74	-0.06	-0.02	1.53	0.01	0.02	1.05	0.00	0.00	0.21
54	0.00	0.00	0.29	-0.04	-0.02	-1.41	0.00	0.02	1.52	0.00	0.00	-0.21
55	0.00	0.00	0.42	-0.04	-0.03	-1.50	0.01	0.02	-0.50	0.00	0.00	0.02
...												
74	0.00	0.00	0.01	-0.03	-0.03	0.15	0.00	0.01	0.00	-0.01	-0.01	0.04
75	0.00	0.00	-0.02	-0.09	-0.09	0.15	0.00	0.02	-0.02	0.00	0.00	-0.07
76	-0.02	-0.02	-0.05									
	RMSE for Z:		16.35203	RMSE for Z:		13.06999	RMSE for Z:		12.7729	RMSE for Z:		19.23435

Figure 8: Cross-validation results

In a subsequent stage, efforts have been made to improve the solution by the development of some simplified neural models. Thus, there have been considered the networks of the three-layer perceptron type, with 6 neurons in the input layer, corresponding to position (X, Y, Z, ψ, θ,φ) of the effector, and a single neuron in the output layer, corresponding to the coordinate of the joint q_3, which determines the Z coordinate. There has been applied the validation technique of the models, the 4-fold cross-validation. The training has been completed using the set of the examples shown in table 2; a remark that should be mentioned in the case being that only the values of the coordinate q_{3j} have been taken into account as output. This set has been divided into four subsets according to the rules described by sets M_k, k=1, ..., 4. For each of the four combinations of the sets M_k, k=1, ...,4, there have been realized four trainings of the network. Thus, during the training, there have been obtained mean square errors that belong to the interval $[10^{-10}, 10^{-1}]$. For each of the four rounds of training-validation corresponding to the four combinations of the sets M_k, k=1, ...,4, there has been determined the root mean square error (RMSE) of approximation of coordinate Z. This has been calculated as an overall mean of all the individual errors $Z_{j,l}^{(k)}$ for a given value k data, where l=1, ..., 4 corresponds to the four neural models obtained within each round. These values are shown in table of Figure 9.

Number of testing points j	Cross-validation															
	M_1				M_2				M_3				M_4			
0	$\Delta Z_{j,1}^{(1)}$	$\Delta Z_{j,2}^{(1)}$	$\Delta Z_{j,3}^{(1)}$	$\Delta Z_{j,4}^{(1)}$	$\Delta Z_{j,1}^{(2)}$	$\Delta Z_{j,2}^{(2)}$	$\Delta Z_{j,3}^{(2)}$	$\Delta Z_{j,4}^{(2)}$	$\Delta Z_{j,1}^{(3)}$	$\Delta Z_{j,2}^{(3)}$	$\Delta Z_{j,3}^{(3)}$	$\Delta Z_{j,4}^{(3)}$	$\Delta Z_{j,1}^{(4)}$	$\Delta Z_{j,2}^{(4)}$	$\Delta Z_{j,3}^{(4)}$	$\Delta Z_{j,4}^{(4)}$
1	-0,001	-0,041	-0,234	0,079	0,000	-0,119	0,461	-0,059	2,256	-1,048	-0,001	-1,409	1,060	0,013	0,064	-0,032
2	-0,001	-0,031	-0,216	0,077	0,000	-0,118	0,428	-0,058	2,166	-0,970	-0,001	-1,375	1,654	0,013	0,068	-0,030
3	-0,001	-0,022	-0,198	0,074	0,000	-0,116	0,402	-0,057	2,091	-0,899	-0,001	-1,343	2,211	0,013	0,071	-0,028
…																
22	0,000	-0,026	0,034	-0,267	0,000	0,044	-3,460	-0,039	1,740	-0,927	0,002	-0,369	-9,481	0,008	-0,029	0,001
23	-0,001	-0,030	-0,032	-0,461	0,000	0,572	-12,696	-0,038	1,555	-3,104	0,000	2,796	10,090	0,010	0,271	0,002
24	-0,002	-0,034	-0,531	-0,024	0,001	1,719	-1,811	-0,037	0,055	-9,585	-0,035	1,152	48,128	0,016	0,578	0,000
25	0,000	-0,038	-1,434	1,673	-0,102	-2,746	59,249	-0,093	59,838	19,992	-2,327	5,756	101,070	15,183	94,398	4,899
26	-14,981	-99,853	-99,944	-46,647	0,102	-111,156	-51,178	12,837	3,938	-20,809	7,291	22,060	-45,740	0,144	-0,949	0,229
27	-0,002	0,040	1,127	-1,091	0,000	8,894	3,049	-0,057	2,046	6,082	0,017	-5,243	-7,538	0,002	0,365	-0,004
28	0,000	0,036	0,562	0,356	0,000	-2,918	12,609	-0,056	1,783	1,850	0,000	-2,480	11,334	0,006	0,367	-0,004
29	0,000	0,032	0,213	0,367	0,000	-1,146	5,186	-0,055	1,587	0,354	-0,002	1,655	15,798	0,007	0,166	-0,003
30	0,000	0,029	0,013	0,185	0,000	-0,177	-2,233	-0,054	1,406	0,024	-0,002	2,654	12,724	0,006	0,090	-0,002
…																
48	0,000	0,017	-0,141	-0,089	0,000	-1,251	1,780	-0,027	0,324	0,283	0,011	-7,329	3,404	-0,003	-0,057	-0,004
49	0,000	0,021	0,127	-0,301	0,000	-1,382	13,452	0,041	0,418	8,121	-0,083	-13,537	-12,827	-0,008	0,566	-0,004
50	-0,001	0,024	1,079	-1,669	0,000	-4,614	-11,118	-0,904	2,237	-12,871	1,757	23,662	-44,820	-4,583	-100,176	0,705
51	-12,486	-178,302	-100,256	-116,688	170,280	103,936	99,415	156,207	112,027	39,528	23,061	-6,690	107,787	-0,221	-0,524	-0,688
52	0,005	-0,036	-1,026	2,379	0,001	10,627	10,384	1,315	-4,863	-5,776	-0,240	10,641	58,543	0,004	0,097	0,007
53	0,000	-0,031	-0,099	-0,055	0,001	-0,721	-16,056	-0,188	-0,233	2,797	-0,010	7,917	22,864	-0,003	0,104	0,007
54	0,000	-0,026	0,160	-0,262	0,001	-0,424	-6,344	-0,084	-0,005	4,455	-0,005	2,124	3,487	-0,006	0,100	0,006
…																
73	0,000	0,037	0,174	-0,320	0,000	-0,256	-0,582	-0,056	-0,489	1,381	-0,001	-0,517	3,082	-0,020	0,077	-0,006
74	0,000	0,029	0,170	-0,323	0,000	-0,259	-0,588	-0,057	-0,556	1,766	-0,001	-0,465	3,433	-0,021	0,078	-0,008
75	0,000	0,019	0,164	-0,325	0,000	-0,262	-0,593	-0,058	-0,643	2,236	-0,001	-0,404	3,745	-0,022	0,081	-0,009
76	-8,3E-05	0,0062	0,155032	-0,32785												
	RMSE=16,01783				RMSE=17,73955				RMSE=8,455625				RMSE=13,45884			

Figure 9: The overall RMSE

The analysis of the data in tables of Figures 8 and 9 shows that there is no significant difference between the errors of approximation of coordinate Z in the case of the two modelling approaches. As a conclusion of this cross-validation, it has been remarked that the method can be applied, but in order to avoid the obstacle it is necessary to envelope it (to cover the obstacle with a smoother surface). The following case studies play the role of evaluating the avoidance method by using the neural models in which the obstacle is enveloped.

3 Case study for an enveloped obstacle

In order to analyse the avoidance precision, based on the same number of training points, the obstacle with three types of envelopes (Figure 10) has been considered.

The point O_1 has the coordinates (500,500,200), while the point O_1' has the coordinates (500,500,450), the distance O_1O_1' being greater than the segment AA'. In order to reduce the number of the training examples, it is considered that ψ=θ=φ=0.

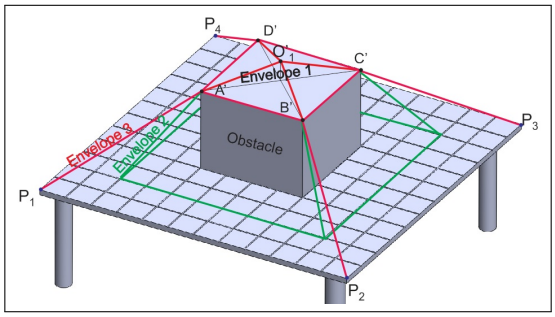


Figure 10: The obstacle envelopes

Training points	Effector coordinates							Joint coordinates ($q_{i,j}$, i=1,...,6; j=1,...,169)					
j	X_j	Y_j	Z_j	Z'_j	ψ_j	θ_j	φ_j	$q_{1,j}$	$q_{2,j}$	$q_{3,j}$	$q_{4,j}$	$q_{5,j}$	$q_{6,j}$
0	1	2	3	4	5	6	7	8	9	10	11	12	13
1	200	200	200	200	0	0	0	20	20	20	20	20	20
2	250	200	200	200	0	0	0	25	20	20	20	20	20
3	300	200	200	200	0	0	0	30	20	20	20	20	20
4	350	200	200	200	0	0	0	35	20	20	20	20	20
5	400	200	200	200	0	0	0	40	20	20	20	20	20
6	450	200	200	200	0	0	0	45	20	20	20	20	20
7	500	200	200	200	0	0	0	50	20	20	20	20	20
8	550	200	200	200	0	0	0	55	20	20	20	20	20
9	600	200	200	200	0	0	0	60	20	20	20	20	20
10	650	200	200	200	0	0	0	65	20	20	20	20	20
11	700	200	200	200	0	0	0	70	20	20	20	20	20
12	750	200	200	200	0	0	0	75	20	20	20	20	20
13	800	200	200	200	0	0	0	80	20	20	20	20	20
14	200	250	200	200	0	0	0	20	25	20	20	20	20
...													
78	800	450	200	200	0	0	0	80	45	20	20	20	20
79	200	500	200	200	0	0	0	20	50	20	20	20	20
80	250	500	200	200	0	0	0	25	50	20	20	20	20
81	300	500	200	200	0	0	0	30	50	20	20	20	20
82	350	500	200	200	0	0	0	35	50	20	20	20	20
83	400	500	200	400	0	0	0	40	50	40	20	20	20
84	450	500	200	425	0	0	0	45	50	42.5	20	20	20
85	500	500	200	450	0	0	0	50	50	45	20	20	20
86	550	500	200	425	0	0	0	55	50	42.5	20	20	20
87	600	500	200	400	0	0	0	60	50	40	20	20	20
88	650	500	200	200	0	0	0	65	50	20	20	20	20
89	700	500	200	200	0	0	0	70	50	20	20	20	20
90	750	500	200	200	0	0	0	75	50	20	20	20	20
91	800	500	200	200	0	0	0	80	50	20	20	20	20
92	200	550	200	200	0	0	0	20	55	20	20	20	20
...													
168	750	800	200	200	0	0	0	75	80	20	20	20	20
169	800	800	200	200	0	0	0	80	80	20	20	20	20

Figure 11: Training data for Envelope 1

For the points in the work plane that do not belong to the obstacle proper or to its envelope, the coordinates of the joints are calculated based on the coordinates (X_j, Y_j, Z_j, ψ_j, θ_j, φ_j)

which define these points.

For the other points in the work plane which belong to the obstacle, or to its envelope, namely those which are at the intersection between the work plane and the envelope of the obstacle, the coordinates of the joints are calculated based on the coordinates of some corresponding points situated on the envelope.

For each of the three envelopes of the obstacle, there has been created and trained a neural network. Table of Figure 11 shows how the training data for Envelope 1 has been achieved.

The training of the neural network has been completed having as input signals coordinates X_j, Y_j, Z_j, ψ_j, θ_j, φ_j of the points (columns 1-3 and 5-7 in table of Figure 11), and as output signals, the coordinates of the joints $q_{i,j}$ (columns 8-13 in table of Figure 8) corresponding to points X_j,Y_j and Z'_j. The same has been applied in the case of Envelopes 2 and 3.

Testing points j	End-effector coordinates									Erors [%]								
	X_j	Y_j	Z_j	Envelope 1 Z'_j	Envelope 2 Z'_j	Envelope 3 Z'_j	ψ_j	θ_j	φ_j	Envelope 1			Envelope 2			Envelope 3		
										$\%Xr_j$	$\%Yr_j$	$\%Zr_j$	$\%Xr_j$	$\%Yr_j$	$\%Zr_j$	$\%Xr_j$	$\%Yr_j$	$\%Zr_j$
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	200	200	200	200	200	200	0	0	0	0.82	1.225	0.73	-0.53	-0.58	-0.87	0.635	0.71	0.24
2	250	250	200	200	200	250	0	0	0	0.24	-0.44	3.06	1.78	1.51	1.72	-0.40	-0.57	-1.09
3	275	275	200	200	200	275	0	0	0	1.54	0.61	2.88	1.18	0.98	3.43	-0.22	-0.33	-1.42
4	300	300	200	200	200	300	0	0	0	1.76	0.76	1.48	0.76	2.25	0.69	-0.05	-0.09	-1.99
5	325	325	200	200	200	325	0	0	0	0.98	0.10	1.61	-1.11	-1.11	6.18	-0.10	-0.15	-2.13
6	350	350	200	200	200	350	0	0	0	0.26	-0.48	2.36	-0.25	0.39	7.94	-0.09	-0.15	-0.60
7	375	375	200	200	300	375	0	0	0	0.06	-0.64	40.21	0.79	-0.59	5.00	0.04	0.06	-0.98
8	400	400	200	400	400	400	0	0	0	0.49	-0.28	-3.94	1.62	-0.39	-2.44	-0.23	-0.23	-0.89
9	425	425	200	412.5	412.5	412.5	0	0	0	1.03	0.14	-2.11	1.70	-0.04	-3.12	-0.22	-0.24	-1.13
10	450	450	200	425	425	425	0	0	0	0.98	0.15	2.17	1.28	0.59	-1.02	-0.02	0.02	-0.81
11	475	475	200	437.5	437.5	437.5	0	0	0	0.41	-0.16	3.75	0.72	1.30	0.43	-0.01	-0.02	-1.70
12	500	500	200	450	450	450	0	0	0	0.38	0.04	-0.71	0.48	1.31	-0.33	0.03	-0.01	-2.82
13	525	525	200	437.5	437.5	437.5	0	0	0	0.61	0.36	-1.22	0.23	0.88	1.10	0.01	0.05	-1.62
14	550	550	200	425	425	425	0	0	0	0.47	0.15	-0.49	0.23	0.33	0.29	-0.19	-0.14	-0.19
15	575	575	200	412.5	412.5	412.5	0	0	0	-0.06	-0.32	-1.54	0.43	0.29	-2.98	-0.04	0.00	-1.04
16	600	600	200	400	400	400	0	0	0	-0.04	-0.28	-5.27	0.55	0.50	-3.98	-0.08	-0.08	-2.81
17	625	625	200	200	300	375	0	0	0	0.38	0.04	24.69	0.47	0.36	3.32	-0.12	-0.13	-2.80
18	650	650	200	200	200	350	0	0	0	0.43	0.03	0.82	0.17	0.35	6.38	-0.26	-0.30	-2.48
19	675	675	200	200	200	325	0	0	0	0.23	-0.18	0.93	-0.41	0.22	3.40	-0.09	-0.13	-2.06
20	700	700	200	200	200	300	0	0	0	0.20	-0.28	0.96	-0.23	0.60	2.62	0.10	0.10	-0.37
21	725	725	200	200	200	275	0	0	0	0.20	-0.38	1.44	-0.46	0.60	-0.05	-0.09	-0.13	2.45
22	750	750	200	200	200	250	0	0	0	0.18	-0.44	0.63	0.14	1.18	4.58	-0.14	-0.19	-0.26
23	800	800	200	200	200	200	0	0	0	0.18	-0.09	0.44	-0.13	-0.92	-0.58	-0.06	-0.07	0.62

Figure 12: Errors for Envelope 1, 2, 3

In order to validate the method and the neural models, there have been considered as input data the coordinates of points (X_j, Y_j, Z_j, ψ_j, θ_j, φ_j) corresponding to the move of the effector on the direct trajectory P_1P_3 (columns 1-3 and 7-9 in table of Figure 12). Based on the coordinates of the joints $q_{i,j}$ simulated by the neural models corresponding to the envelopes, by means of relations (1)-(6), there have been calculated the effector coordinates and they have been graphically represented in Figure 13. The error obtained through the simulation on the neural model corresponding to each envelope is shown in table of Figure 12 (columns 10 - 18).

The results in table of Figure 12 show that Envelope 1 does not solve the problem at the borders of the intersection of the direct trajectory with the obstacle, as it is possible for the latter to be hit. This problem is solved in the case of Envelopes 2 and 3. The precision obtained by simulation for the models Envelope 2 and Envelope 3 can be accepted only in the case of some handling applications that do not require a high level of precision. This precision can be improved by increasing the number of training examples and by increasing their density in the workspace. In this research, there have been used only 169 training examples for the entire

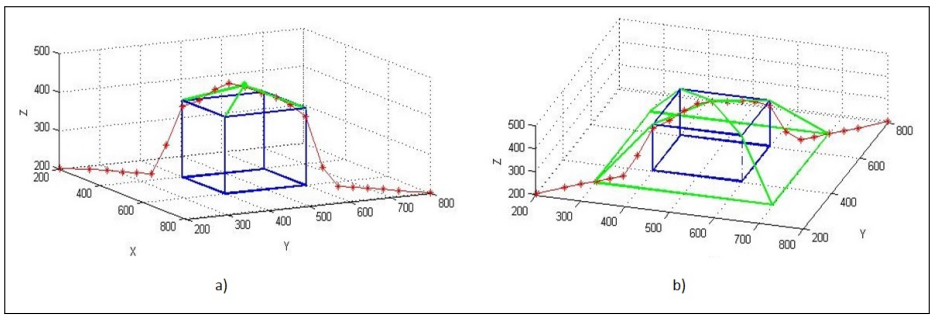


Figure 13: Effector coordinates for a. Envelope 1; b. Envelope 2

workspace, the distance between two successive points being 50 mm.

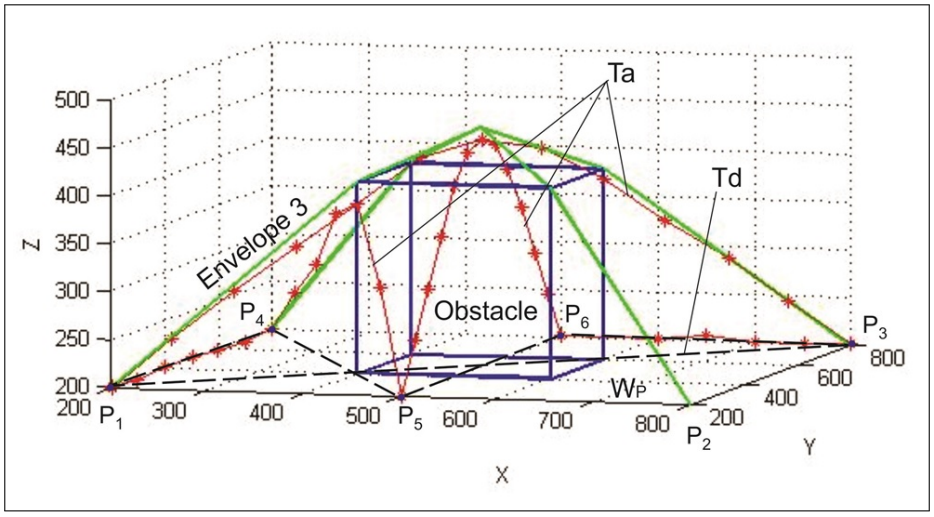


Figure 14: Direct and simulated trajectory

In order to test the obstacle avoidance in the case of programming several trajectories in the work plane (Envelope 3), there has been considered the robot's move on direct trajectories under the form of straight segments between points P_1-P_4-P_5-P_6-P_3-P_1 (Figure 14). The points P_1, P_4, P_5, P_6, P_3 are points in which the robot has to complete operations and where the effector has to position itself with an accuracy corresponding to the application. All the points P_1, ..., P_6 are situated in the work plane Wp. For the validation, there has been used the neural network for Envelope 3 and the results are shown in Figure 14.

The analysis of the results in Figure 14 reveals that, although the input data consisted of coordinates of some points on the direct trajectories Td, the neural model offers the joint coordinates that make the robot avoid obstacle when direct trajectory Td intersects the envelope of this obstacle. When the direct trajectory does not intersect the obstacle, the neural model NM provides data that leads the robot's effector on the direct trajectory Td.

4 Conclusions and future work

In order to validate the method by numerical research, there has been considered a robot with six degrees of freedom that has to move between two points in the workspace Wp. In plane Wp, there has been considered a parallelepiped-shaped object which has to be avoided. The research aimed to obtain several neural models of the robots kinematics based on a certain

method of creation of the set of training examples. Within the research, it has been noticed that the model obtained by the training of the robot so that it move on a trajectory that avoids, to the limit, the straight trajectory (the completed trajectory follows the obstacles outline) does not approximate well the move along the axis OZ at the frontier of the obstacle. There have been recorded relative errors of approximately 50% in such points.

In order to obtain an improved model, there have been considered three envelopes that dress the obstacle. For each envelope, there has been trained a neural network, each having the same number of training examples. The set of training examples has been generated by inverse kinematics analysis, considering a cloud of equally distanced points in the robot's work plane. The joint coordinates have been generated depending on the obstacle's envelope, in accordance with rule R described in the paper.

It has been noted that in all the three cases the neural model provides the joint coordinates that lead the end-effector on a bypass trajectory. In all the three cases studied, the bypass trajectory intersects the obstacle near the points that limit its superior base ($A'B'C'D'$). The positive deviation from the coordinate Z does not affect the obstacle avoidance, only the negative ones.

The analysis of the errors in the case of the coordinates (Xr_j, Yr_j, Zr_j) simulated by the neural network relating to the programmed coordinates (X_j, Y_j, Z_j) shows that the results are influenced by the choice of the obstacle envelope as follows:

- for Envelope 1:
  - the coordinates Xr_j and Yr_j are approximated with errors less than 2%
  - there are important errors in the case of coordinates Zr_j simulated by the neural model in the area close to the obstacle (40%, table of Figure 12, row 7)
  - close to points A' and C', the deviant trajectory intersects the obstacle ($\%Zr_j$ = -5.27%, table of Figure 12, row 16)
- for Envelope 2:
  - the coordinates Xr_j and Yr_j are approximated with errors less than 2.5%
  - the coordinates Zr_j simulated by the neural model are quite well approximated ($\%Zr_j$<8%, table of Figure 12, row 6)
  - close to points A' and C', the deviant trajectory intersects the obstacle ($\%Zr_j$= -3.98%, table of Figure 12, row 16)
- for Envelope 3:
  - the coordinates Xr_j and Yr_j are approximated with errors less than 1%
  - the coordinates Zr_j simulated by the neural network are well approximated ($\%Zr_j$<3%)
  - close to points A' and C', the deviant trajectory intersects the obstacle ($\%Zr_j$=-2.81%, table of Figure 12, row 16).

The problem of the intersection between the bypass trajectory and the obstacle can be solved by the choice of an envelope that exceeds the obstacle in all its points. Thus, for the case in point, if one chooses the parallelepiped $ABCDA'B'C'D'$ with sides 10% larger that the dimensions of the obstacle, the problem of the collision is going to be avoided. Another option is to increase the training examples number.

Future research aims to optimize the avoidance trajectories and to improve the position accuracy in the working points. The trajectory optimization will account for accuracy and

energy consumption refinements. The positioning accuracy in the working points will be made by choosing a denser cloud of points around them. Further research aims to study the method presented in this paper using other types of neural networks and make a comparative analysis of the results.

Bibliography

[1] Ciupan E.; Bojan I.(2009); Own Software Used in the Modelling of an Economic Supplying Activity, *Proceeding of the 5th International Conference on the Management of Technological Changes*, ISSN 1726-9679, (1): 209-212.

[2] Ciupan E. (2007); Software Designed for Modelling and Simulating Using Three-layer Neural Networks, *Annals of DAAAM 2008 & Proc. of the 19th International DAAAM Symposium*, ISBN 978-960-89832-7-4, (1): 275-276.

[3] Ciupan E. (2010); A Model for the Kinematical Analysis of a Six Degrees of Freedom Parallel Robot, *Proceedings of the 2010 IEEE International Conference on Automation, Quality and Testing, Robotics AQTR 2010*, ISBN 978-1-4244-6724-2, (1): 261-264.

[4] Ciupan E. (2012); Modelling a Handling Robot to Avoid an Obstacle, *Proc. of the 2012 IEEE International Conference on Automation, Quality and Testing, Robotics AQTR 2012*, ISBN 978-1-4673-0702-4, (1):391-395.

[5] Cong S.; Liang Y. (2009); PID-Like Neural Network Nonlinear Adaptive Control for Uncertain Multivariable Motion Control Systems, *IEEE Trans. Ind. Electron*, ISSN 0278-0046, 56, (10): 3872-3879.

[6] Debbache G.; Bennia A.; Gola N. (2007); Neural Networks-Based Adaptive State Feedback Control of Robot Manipulators, *International Journal of Computers Communications & Control*, ISSN 1841-9836, 2(4):328-339.

[7] Feng Y.; Wanh Y.; Yang Y. (2012); Inverse Kinematics Solution for Robot Manipulator Based on Neural Network under Joint Subspace, *International Journal of Computers, Communications & Control*, ISSN 1841-9836, 7(3): 459-472

[8] Karray F.; Gueaieb W.; Al-Sharhan S. (2005); The Hierarchical Expert Tuning of PID Controllers Using Tools of Soft Computing, *IEEE Transactions on Systems, Man, and Cybernetics*, Part B , 35(6):1283-1294

[9] Lee S. H. et al. (2007); Patent KR100752098, Robot System Based on Neural Network.

[10] Lewis F. L.; Jagannathan S.; Yesildirak A. (1998); *Neural Network Control for Robots Manipulators and Nonlinear Systems*, ISBN0748405968, Tylor & Francisc Inc. Bristol, 1998.

[11] Mann G. K. I.; Hu B-G.; Gosine R.G. (2001); Two-Level Tuning of Fuzzy PID Controllers, *IEEE Transactions on Systems, Man, and Cybernetics*, Part B, 31(2): 263-269.

[12] Min T. et al. (2012); Patent CN102346489, Pulse Neural Network Based Method for Controlling Object Tracking of Robot.

[13] Park K. T. et al. (2005); Patent KR20070072314, Method for Controlling Pose of Robot With Using Neural Network, Recording Medium Thereof, Apparatus for Controlling Pose of Robot With Using Neuron-Network and Robot Therewith.

[14] Pisla D. L.; Itul T. P.; Pisla A. (2007); Considerations Regarding the Geometrical Modelling of Parallel Mini-Manipulators With Triangle Platform, *Proc. of 11th Int. Research/Expert Conference Trends in the Development of Machinery and Associated Technology TMT 2007*, ISBN 9958-617-30-7: 607-610.

[15] Rigatos G. (2009); Model-based and model-free control of flexible-link robots: A comparison between representative methods, *Applied Mathematical Modelling*, 33(10): 3906-3925.

[16] Siciliano B.; Katib O. (2008); Springer Handbook of robotics, Springer-Verlag, Berlin Heidelberg, ISBN 978-3-540-23957-4.

[17] Uang H. J.; Lien C.C. (2006); Mixed H2/H & PID tracking control design for uncertain spacecraft systems using a cerebellar model articulation controller, *IEEE Proc.- Control Theory and Applications*, 3(1): 1-13.

[18] Yu D. L.; Chang T. K.; Yu D.W. (2005); Fault Tolerant Control of Multivariable Processes Using Auto-Tuning PID Controller, *IEEE Transactions on Systems, Man, and Cybernetics*, Part B , 35 (1): 32-43.

[19] Yu W.; Rosen J. (2013); Neural PID Control of Robot Manipulators application to an Upper Limb Exoskeleton, *IEEE Transactions on Cybernetics*, 43(2): 673-684.

[20] Zhang P.Y.; Lu T.S.; Song L.B. (2004); RBF networks-based inverse kinematics of 6R manipulator, *International Journal of advanced manufacturing technology*, 26(1): 144-147.

[21] Zilouchian A.; Jamshidi M. (2001); Intelligent Control Systems using Soft Computing Methodologies, CRC Press LLC, ISBN 0-8493-1875-0.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 9(5):555-569, October, 2014.

A Fairness Load Balancing Algorithm in HWN Using a Multihoming Strategy

Y. Donoso, C. Lozano-Garzon, M. Camelo, P. Vila

Yezid Donoso
Universidad de los Andes
Bogotá, Colombia, South America
ydonoso@uniandes.edu.co

Carlos Lozano-Garzon
Universidad de los Andes
Bogotá, Colombia, South America
& Universitat de Girona, Girona, Spain.
calozanog@ieee.org

Miguel Camelo, Pere Vila
Universitat de Girona, Girona, Spain.
miguel.camelo@udg.edu, pere.vila@udg.edu

Abstract: Due to the growth of the number of intelligent devices and the broadband requirements, between others technical requirements, of the new applications, suppose a new challenge in planning, maintenance and resource allocation in mobile networks for the telecommunication operators. Service providers must ensure a quality of service for users in a new environment based in Heterogeneous Wireless Networks (HWN). A good way to achieve this goal is to prevent the quantity of services of each mobile users being connected to the same access networks and therefore reducing the possibility of overloading it. This paper presents a load balancing optimization scheme that enables operators to make decisions about re-allocation of each of the services in different access networks, keeping the required Quality of Service (QoS). In this paper, we propose 1) a mathematical model addressed as a fairness resource allocation in order to obtain a global load balancing, and 2) a two-step algorithm based on the anchor-adjustment heuristic to solve it. Our algorithm contribute to unload the network with maximum load while at the same time, the other networks are balanced. As a result, we show that our algorithm finds (near)-optimal solutions while keeps low complexity.
Keywords: Fairness, load balancing, multihoming, quality of service, heterogeneous wireless networks (HWN).

1 Introduction

Nowadays, mobile operators have a great challenge in planning, maintenance and optimization of their complex network infrastructure [2]. Many of these challenges are related to the continuous growth of both mobile subscriptions and mobile traffic; in [3] the Global mobile Suppliers Association estimates that there are around 1.3 billion of broadband mobile subscriptions in the second quarter of 2013, and that the monthly traffic was around 885 Petabyte in 2012 and is expected to reach to 11.2 Exabyte in 2017 [4]. This growing, together with the need of allowing a quality of service that ensures connectivity and mobility to its users, regardless of the access technology, constitute the main topic of our study.

Current infrastructure deployed by the majority of mobile operators is made up by a combination of radio access technologies (RAT). It is possible that during the network operation of the network some of these access channels could be overloaded due to a sudden growth of traffic. Hence it is necessary to perform a good management of the resource allocation, specifically bandwidth in this case, over all the access infrastructures. The final aim is to ensure a balance between the use of network resources and the number of user connections; likewise, operators must allow user mobility between technologies without users perceiving the change [5]. Thus, it is essential for operators to use decision-making algorithms to manage both their network resources and the connection of the services. This decision could be based on the QoS requirements, among other parameters, and it can be seen as a RAT selection process.

Combining the advantage of the existence of user equipment and network protocols to provide connections through multiple interfaces to different kind of networks [10], it would provide network benefits such as ubiquitous access, reliability, load balancing, among others [6]. Specifically, multihoming could separate a flow between multiple points of attachment (simultaneously active or not) of a node, usually by choosing the less loaded connection or according to preferences on the mapping between flows and interfaces. Following the latest idea, we initiated our study about the Always Best Connected (ABC) problem in HWN with an approach based on the possibility that the mobile user could make the decisions about which network it wants to be connected [7]. We designed a Vertical Handover (VHO) Decision Algorithm that allows the user terminal to start a proactive re-allocation of the mobile based on parameters such as: user preferences, QoS requirements, and network conditions. This process has as aim to avoid the over burdening of any interface.

In this paper, we propose an efficient decision-making algorithm to perform a load balancing by re-allocation of connected services among networks by using multihoming as main strategy. However, the decisions are made on the operator side by using both local and global information. The proposed algorithm is based on the anchor-adjustment heuristic proposed by Tversky and Kahneman in [8], which runs in quadratic time proportional to the number of mobiles in the network.

This paper extends the conference paper [1]. The key additions of this journal version are as follows. First, Section 2 describes several works about the use of multihoming in this research scenario. Second, this paper contains an additional explanation of the proposed model in Section 3 and the algorithm designed in Section 4. Finally, this paper contains an additional scenario composed by seven networks and a variable number of mobiles; our obtained results are contrasted with the results of round robin and least connected algorithms, two ore the most used load balancing algorithms.

The remainder of this paper is organized as follows. In Section 2 we present the related work on load balancing in cellular networks using multihoming as main strategy. In Section 3 we introduce the mathematical model that encode the objective function to obtain a global load balancing among HWN under QoS constraints. In section 4 the load balancing algorithm based on the anchor-adjustment heuristic is presented. The experimental results about the performance of our proposal compared to round robin and least connected algorithms are shown in Section 5. Finally, concluding remarks and directions for further research are given in Section 6.

2 Related Work

Multihoming and Load Balancing strategies have been topics of several research projects. A mathematical model to load balancing is presented in [9]. This model aims to minimize the load of network by re-allocating services from the more loaded network. Sousa, Pentikousis and Curado present in [10] the architectural goals and system design principles for multihoming, and review

different approaches. In addition, they show in a survey, how multihoming is supported at the different levels of the OSI layers, covering all recent proposals based on a locator/identifier split approach. A mathematical model for heterogeneous network and its performance, considering both multihoming and network coding, is presented in [11]. They propose an optimal resource allocation by deciding which data should be transmitted on which interface and, simultaneously, which coding parameters should be used on the data. They showed that the combination of multihoming and network coding can improve significantly the service rate of Access Points (WiFi) and Base Stations (Cellular System) and reduce the system delay. An alternative technique based on traffic splitting using common radio resource management for Long Term Evolution (LTE) and High-Speed Downlink Packet Access (HSDPA) networks is presented in [12]. They propose a mathematical model to solve the traffic split problem. In their research, they found that to maximize the throughput it is necessary minimize the transmission delay between both networks. Then, the split ratios can be dynamically adjusted according to the channel qualities and the load status of the networks.

Some other research studies about the use of multihoming in the cellular heterogeneous networks are related with the VHO process. In [13], authors present a transport-layer scheme to support VHO between Universal Mobile Telecommunications System (UMTS) and Wireless Local Area Network (WLAN) using Stream Control Transmission Protocol (SCTP). Their method is based in the multihoming capability and the dynamic address reconfiguration extension of SCTP; through it, they obtain a decrease in handover delay and improve throughput performance. The study of Liu, Boukhatem, Martins and Bertin in [14] propose a multihoming approach to implement a seamless VHO for UMTS and WiMAX over integrated and tight coupling architectures. Based on these architectures, authors design a sublayer of OSI layer 2 to be added to the Radio Network Controller (RNC) and Mobile Station (MS); this sublayer implements a dual retransmission queue scheme to enable a soft handover that can eliminate packet losses and reduce handover latency significantly.

Paik et al. [15] also designed a seamless VHO mechanism using multihoming for mobile networks. They implement a multihomed mobile access point, which was tested over heterogeneous networks formed by Wireless Broadband Network (WiBro) and HSDPA access technologies. As a result, they obtained a significant reduction of handover latency by reducing IP connection latency. Authors in [16] also use the seamless handover and multihoming techniques but their aim was to increase the access availability, which is fundamental for QoS and critical services. They propose one method to calculate and distribute network topology information with estimated availability, which allows to predict the overall availability of accessible networks. This is a criteria, among others, used for a handover decision in the IMH 802.21 framework. To test the proposal they used three different access technologies, WLAN, WiMAX and UMTS.

Finally, other researchers have studied the use of multihoming as strategy to make a better distribution of the bandwidth charge. Sungwook and Varshney [17] worked on a dynamic on-line bandwidth reservation algorithm for some multimedia services over cellular networks. This algorithm was designed to control bandwidth according to the priority of traffic services and current network traffic conditions. Later, in [18] they proposed another adaptive on-line algorithm for the multimedia services but it is based on the minimization of the maximum available bandwidth in each cell in order to keep the load balancing. They compared the performance of both proposed schemes with the ABR scheme and the CAC provision scheme, obtaining an appropriate performance balance between contradictory requirements.

In this paper, we propose both a generic mathematical model to load balancing in HWN, which uses as a main objective function the concept of *fairness* over the networks, and a scalable two-step algorithm with low computational complexity that can be implemented by using VHO.

3 A Mathematical Model for Load Balancing

Since the resources are limited, an incorrect allocation could impact both the performance of the network and the satisfaction perception of the users. Therefore, it is necessary to define a mathematical model that encodes the requirements of the user, the environment constraints and the main target: to balance the load among different networks while ensuring the QoS requirements are met. In the following sections, we define the variables, functions and parameters of the mathematical model, and a solution that gives us an optimal distribution of loads across multiple networks under QoS constraints.

3.1 Network Load

Let N, M and S be the sets of n networks, m mobiles and s services that compose a Cellular System, respectively (See Figure 1). Additionally, let $y_{j,k} \in [0, 1]$ be a binary parameter that indicates if the service k of the mobile j is activated or not. We calculate the load of the network i (α_i) as the sum of demanded bandwidth (D_k) of each connected service (k), for each mobile (j) over the total capacity of the network channel (C_i).

$$\alpha_i = \frac{\sum_{j=1}^{m} \sum_{k=1}^{s} D_k \cdot x_{i,j}^k \cdot y_{j,k}}{C_i}, \forall i \tag{1}$$

Where $x_{i,j}^k = 1$ if the service k of the mobile j is connected to the network i, or 0 otherwise.

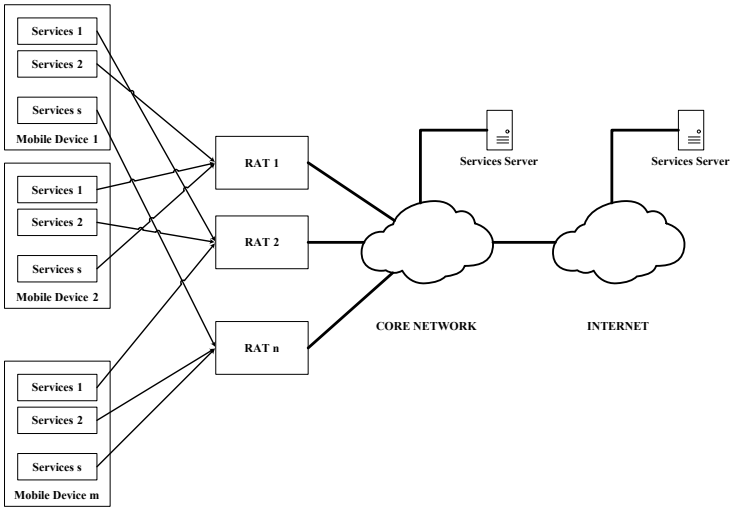


Figure 1: Multihoming Cellular System

3.2 Load Balancing: Jain's Index Functions

Fairness is a concept related with "equality" in the resource allocation widely used in many research fields, including those in Wireless Networks [19]. The Jain's Index function [20] can be formulated to measure the fairness of resource allocation among networks, i.e. it can give us an accurate measurement of how well-distributed is the network load in the Cellular System. This function can be formulated as follows:

$$f_1(\alpha) = \frac{[\sum_{i=1}^{n} \alpha_i]^2}{n \cdot \sum_{i=1}^{n} \alpha_i^2}, \alpha_i \geq 0, \forall i \tag{2}$$

$f_1(\alpha)$ is a continuous map in to the real interval [0,1], where a value closer to 0 represent a unbalanced load in the system, whilst a value near to 1 represents a fairer allocation.

3.3 Constraints

In this model we need to consider several constraints in order to ensure the adjustment to the real-life networks. Some constraints are related to QoS requirements and others to the service connectivity.

QoS constraints

We consider only two QoS parameters in our modeling: Bandwidth and Signal Strength. However, it is important to remark that the model can be extended by including any other QoS requirements. The bandwidth is a resource that is not unlimited, therefore, it is necessary to ensure that all the services connected to a given network receive the requested bandwidth, otherwise the network must reject the connection. In other words, the service k of the mobile j can be connected to the network i if and only if:

$$D_k \cdot x_{i,j}^k \leq AB_i \quad \forall i \in N, \forall j \in M, \forall k \in S \tag{3}$$

Where AB_i is the available bandwidth of the network i.

The Received Signal Strength Indication (RSSI) is the relative received signal strength in a wireless environment. It measures the power level being received by the antenna, where higher values of the RSSI imply stronger received signals. If this value is below a certain threshold $RSSI_{th}$, we assume that the quality of the communication between the mobile and the base station is very poor. Thus, we can define that the service k of the mobile j can be connected to the network i if and only if:

$$x_{i,j}^k \leq z_{i,j}^k \quad \forall i \in N, \forall j \in M, \forall k \in S \tag{4}$$

Where the parameter $z_{i,j}^k = 1$ if $\frac{RSSI_{i,j}^k}{RSSI_{th}} \geq 1$, or 0 otherwise.

Actived services constraint

This constraint ensures that all activated services of each mobile device j must be connected to some network in N.

$$x_{i,j}^k \leq y_{j,k} \quad \forall i \in N, \forall j \in M, \forall k \in S. \tag{5}$$

Connectivity constraint

Through this constraint we ensure that each service used by the mobile j is connected to only one network.

$$\sum_{i=1}^{n} x_{i,j}^k = \max_{1 \leq i \leq n} \{z_{i,j}^k \cdot y_{j,k}\} \quad \forall j \in M, \forall k \in S. \tag{6}$$

Summary

A summary of the proposed mathematical model is presented as follows:
Maximize

$$f_1(\alpha) = \frac{[\sum_{i=1}^{n} \alpha_i]^2}{n \cdot \sum_{i=1}^{n} \alpha_i^2}, \alpha_i \geq 0 \ (2)$$

Subject to

$$D_k \cdot x_{i,j}^k \leq AB_i, \quad \forall i \in N, \forall j \in M, \forall k \in S \ (3)$$

$$x_{i,j}^k \leq z_{i,j}^k, \quad \forall i \in N, \forall j \in M, \forall k \in S \ (4)$$

$$x_{i,j}^k \leq y_{j,k}, \quad \forall i \in N, \forall j \in M, \forall k \in S \ (5)$$

$$\sum_{i=1}^{n} x_{i,j}^k = \max_{1 \leq i \leq n} \{z_{i,j}^k \cdot y_{j,k}\}, \quad \forall j \in M, \forall k \in S \ (6)$$

$$x_{i,j}^k \in \{0, 1\}, \quad \forall i \in N, \forall j \in M, \forall k \in S \ (7)$$

4 Proposed Algorithms

The resource allocation problem has been proved to be NP-Complete [21], therefore, an efficient algorithm that solves this problem is not known yet. In addition, our model includes an objective function (see equation 2) that can only be solved by Mixed-Integer Nonlinear Fractional Programming methods [22], which also difficult the possibility of finding an optimal solution even in small problems. Thus, it is necessary to implement heuristics that either solve the problem quickly or find an approximate solution when classic methods fail.

Based on the anchor-adjustment heuristic proposed in [8], we present a two-step algorithm that solves the above mathematical model in quadratic time proportional to the number of mobiles in the network. The proposed algorithm (see Algorithm 1) finds a (near)-optimal load balancing among networks by performing a re-allocation of connected services. Roughly speaking, the first step of the algorithm tries to balance the load of the network by distributing the services from the more loaded network to the one with minimum load. Second step performs an optimization on the result obtained from the first one by using a load distribution method with local information.

Algorithm 1 Two-Step algorithm based on anchor-adjustment heuristic

Require: List of the set of Available Networks $AN_{j,k} = \{t_1, \ldots, t_p\}, \forall j \in M, \forall k \in S$ and $p \leq n$
Require: Set of actual connection of mobiles and their services $X = \{x_{1,1}^1, \ldots, x_{n,m}^s\}$
Ensure: A (re) allocation of each used service by each connected mobile device.
1: Call Anchor algorithm based on a Max-Min strategy at network level (Algorithm 2)
2: Call Adjustment algorithm based on local information (Algorithm 3)
3: **return** A set of connections of mobiles and their services $X = \{x_{1,1}^1, \ldots, x_{n,m}^s\}$

We assume that for each mobile $j \in M$ and each service $k \in S$ of j, we have access to the set of available networks $AN_{j,k} = \{t_1, \ldots, t_p\}, \forall j \in M, \forall k \in S$ and $p \leq n$, which is derived by selecting those networks that meet both requirements the RSSI threshold and the available bandwidth (see constraints 3 and 4). We will see below that the time complexity of this two-step algorithm is bounded by $O(m^2)$, when $m >> n$ and $m >> s$.

4.1 First Step: Anchor based on a Max-Min strategy at network level

Given the set N of networks and a valid set of connections $X = \{x^1_{1,1}, \ldots, x^s_{n,m}\}$, i.e. set of connections that meet the constraints 5 and 6, the algorithm finds the network $i_{max} \in N$ with maximum load $\alpha_{i_{max}}$ and selects a random service k connected to it. Then, k is moved to the available network with minimum load $i_{min} \in N$ that is available for k. In case that there are no alternative networks to connect s, i.e. either there are not more reachable networks to connect the service or all the networks have the same load, the service is not re-allocated. Note that although the Algorithm 2 can obtain an optimal load balancing among networks in most cases, the addition of constraints into the mathematical model may permit that the algorithm can be stuck in local minimum. For example, if all services connected to $i_{max} \in N$ have not alternative networks to move, then the algorithm can not perform a load balancing among any network. The load balancing is obtained by the assignation of each service to the feasible network with the minimum load at that moment.

Algorithm 2 Anchor based on a Max-Min strategy at network level

Require: List of the set of Available Networks $AN_{j,k} = \{t_1, \ldots, t_p\}, \forall j \in M, \forall k \in S$ and $p \leq n$
Require: Set of actual connections of mobiles and their services $X = \{x^1_{1,1}, \ldots, x^s_{n,m}\}$
Ensure: A (re) allocation of each used service by each connected mobile device.

1: $i_{max} \leftarrow 0$, $i_{min} \leftarrow 0$, $l \leftarrow 0$;
2: **repeat**
3: Compute $\alpha_i, \forall i \in N$;
4: $i_{max} \leftarrow i \mid i \in N$ and $\max_i\{\alpha_i\}$;
5: Select any $x^k_{i,j} = 1 \mid i = i_{max}$;
6: $i_{min} \leftarrow t \mid t \in AN_{j,k}$ and $\min_t\{\alpha_t\}$;
7: **if** $\alpha_{i_{min}} < \alpha_{i_{max}}$ **then**
8: $x^k_{i_{max},j} = 0$, $x^k_{i_{min},j} = 1$;
9: **end if**
10: $l = l + 1$
11: **until** $l = m \cdot s$
12: **return** A set of connections of mobiles and their services $X = \{x^1_{1,1}, \ldots, x^s_{n,m}\}$

The complexity of the Algorithm 2 is $\Theta(m^2 \cdot s^2)$, where m is the number of mobile devices, and s the number of services. Remark that although the main loop has $\Theta(m \cdot s)$ iterations, step 6 of the algorithm implies to find the services connected to a specific network. This operation takes at most $\Theta(m \cdot s)$ steps, i.e. we need to classify each mobile and its services regarding which network they are connected to.

4.2 Second Step: Adjustment based on local information

Algorithm 4.2 uses the result of the Algorithm 2, and tries to improve the allocation of services by using an iterative process of adjustment. The algorithm iterates on each $j \in M$ and moves each one of its service k to its available network with minimum load. In case that there are no alternative networks to connect the service k, i.e. either there are not more reachable networks to connect the service or all the networks have the same load, the service is not re-allocated.

Observe that Algorithm 3 permits to leave local optimal from Algorithm 2 in a greedy way: in each iteration, the service $k \in S$ of each mobile $j \in M$ is moved to its less loaded available network (local information). However, we can not ensure an optimal solution at the end of Algorithm 3 because in each iteration, the actual load of the networks has been influenced by

both the initial allocation of services (from step 1) and the allocation of services performed before the current iteration.

Algorithm 3 Adjustment based on local information

Require: List of the set of Available Networks $AN_{j,k} = \{t_1, \ldots, t_p\}, \forall j \in M$, $\forall k \in S$ and $p \leq n$
Require: Set of actual connection of mobiles and their services $X = \{x^1_{1,1}, \ldots, x^s_{n,m}\}$
Ensure: A (re) allocation of each used service by each connected mobile device.

```
for 1 ≤ j ≤ m do
  for 1 ≤ k ≤ s do
    i_act ← i | x^k_{i,j} = 1;
    i_min ← t | t ∈ AN_{j,k} and min_t{α_t};
    if α_{i_min} < α_{i_act} then
      x^k_{i_act,j} = 0, x^k_{i_min,j} = 1;
    end if
  end for
end for
return A set of connections of mobiles and their services X = {x^1_{1,1}, ..., x^s_{n,m}}
```

We found that the complexity of the Algorithm 3 is $\Theta(m \cdot s)$, where m is the number of mobile devices, and s the number of services. Note that this algorithm has a linear time complexity with respect to the total number of mobiles. In general, the combination of Algorithm 2 and 3 gives us a time complexity equal to $\Theta(m \cdot s + m^2 \cdot s^2)$. However, since the number of mobiles is usually larger than the number of services and networks, the time complexity of the Algorithm 1 has as upper bound $O(m^2)$.

5 Experimental Results

We propose an experimental environment composed for seven different radio access technologies and a finite number of mobile devices, each one with at most three different network interfaces to supporting data, voice, and video services.

Table 1: Access Network Bandwidth

Network	EDGE	HSPA	WiMax	HSPA+	WiFi G	WiFi N	LTE
Bandwidth(Mbps)	0.384	14.4	37.0	42.0	54.0	100	100

Table 2: Requested Bandwidth for Service

ID	Service	Required Bandwidth (Mbps)
1	Voice	0.012
2	Data	0.028
3	Video	0.128

Tables 1 and 2 show the characteristics of each access network and the minimum required bandwidth to ensure QoS for each service. To verify the efficiency and effectiveness of our algorithm, we compare the solutions found by our algorithm with respect to those obtained by solving the mathematical model and by using both the Round Robin and Least Connected algorithms. The mathematical model was implemented in GAMS [23], and it was solved by using the global optimization solver BONMIN (Basic Open-source NonlinearMixed INteger programming) [24].

Two different scenarios to evaluating the proposed algorithm are presented in the following sub-sections.

5.1 Scenario 1: Three networks and 10 mobiles

The first scenario is composed of three networks (WiMax, EDGE and HSPA) and 10 mobile devices, where each one of them has two activated services: data and voice. Mobiles are randomly distributed among networks and the RSSI threshold is set to 10. Figure 2, and Table 3 show the distribution of mobiles among the networks and the RSSI values of each mobile according to a random distribution. Table 4 (a) presents an initial connection matrix, which is the result of a random selection process on the mobile RSSI matrix, which ensures the required bandwidth of each service.

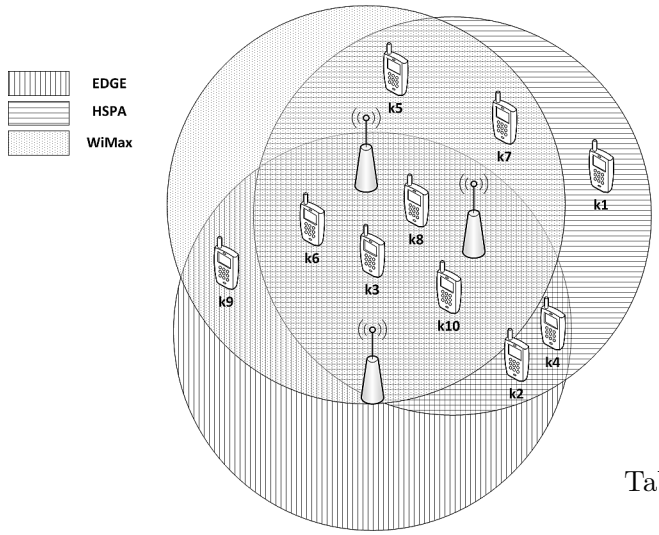


Figure 2: Distribution of mobiles in scenario 1

Mobile	WiMax	EDGE	HSPA
k1	0	3	26
k2	9	27	23
k3	16	19	29
k4	5	15	12
k5	29	4	12
k6	14	17	24
k7	24	8	29
k8	15	14	18
k9	28	15	0
k10	19	21	11

Table 3: RSSI for each mobile in scenario 1

(a)		
Mobile ID	Service 1	Service 2
k1	HSPA	HSPA
k2	HSPA	HSPA
k3	EDGE	EDGE
k4	HSPA	HSPA
k5	HSPA	HSPA
k6	EDGE	EDGE
k7	HSPA	HSPA
k8	EDGE	EDGE
k9	EDGE	EDGE
k10	EDGE	EDGE
Jain's Index = 0.3510		

(b)		
Mobile ID	Service 1	Service 2
k1	HSPA	HSPA
k2	HSPA	HSPA
k3	WiMax	WiMax
k4	HSPA	HSPA
k5	HSPA	HSPA
k6	WiMax	WiMax
k7	HSPA	HSPA
k8	WiMax	WiMax
k9	WiMax	WiMax
k10	WiMax	WiMax
Jain's Index = 0.5586		

(c)		
Mobile ID	Service 1	Service 2
k1	HSPA	HSPA
k2	EDGE	HSPA
k3	WiMax	WiMax
k4	HSPA	HSPA
k5	WiMax	WiMax
k6	WiMax	WiMax
k7	WiMax	WiMax
k8	HSPA	WiMax
k9	WiMax	WiMax
k10	WiMax	WiMax
Jain's Index = 0.6653		

Table 4: Connection of services: (a) Initial, (b) obtained from Algorithm 2, and (c) obtained from Algorithm 3

From the initial connection matrix, the computed Jain's Index (See function (2)) was 0.3510 (See Table 4 (a)). Once we ran the Algorithm 2, this value raised up to 0.5586 (See Table 4 (b)). Finally, by using the connection matrix resulting from Algorithm 2 as input for 3, the Jain's Index increased to 0.6653 (See Table 4 (c)), while the optimal value of such function calculated by BONMIN was 0.7070 in 17 seconds (See Table 5). We emphasize that the solution found by our proposal has only a relative error of 5.9% with respect to the optimal one, and its execution time was less than 1 second.

Mobile ID	Service 1	Service 2
k1	HSPA	HSPA
k2	HSPA	HSPA
k3	WiMax	WiMax
k4	HSPA	HSPA
k5	HSPA	WiMax
k6	WiMax	HSPA
k7	HSPA	HSPA
k8	EDGE	HSPA
k9	WiMax	WiMax
k10	WiMax	WiMax
Jain's Index = 0.7070		

Table 5: Optimal connection of services

5.2 Scenario 2: Seven networks and a variable number of mobiles

In this scenario, we show the behavior of the proposed algorithm when the number of mobiles is increased. The scenario is composed for seven different radio access networks (EDGE, HSPA, WiMax, HSPA+, WiFi G, WiFi N, and LTE), and a set of a distributed randomly mobile devices over those networks, where each one has three activated services and the number of them is increased from 10 to 1000. We also implement both the round Robin (RR) and Least Connected (LC) algorithms [25] to compare their performance with respect to our proposed algorithm. The pseudo-codes of these algorithms are described in Algorithms 4 and 5. It is important to note that our version of RR and LC also include a random selection process when the next network in the list can not be reached by a given mobile.

Algorithm 4 Round Robin Algorithm

Require: List of the set of Available Networks $AN_{j,k} = \{t_1, \ldots, t_p\}, \forall j \in M, \forall k \in S$ and $p \leq n$
Require: Set of actual connection of mobiles and their services $X = \{x^1_{1,1}, \ldots, x^s_{n,m}\}$
Ensure: A (re) allocation of each used service by each connected mobile device.

$i_{new} \leftarrow -1$;
for $1 \leq j \leq m$ **do**
  for $1 \leq k \leq s$ **do**
    $i_{act} \leftarrow i \mid x^k_{i,j} = 1$;
    $i_{new} \leftarrow (i_{new} + 1) \bmod (n-1) + 1$;
    if $t_{i_{new}} \in AN_{j,k}$ **then**
      $x^k_{i_{act},j} = 0$, $x^k_{i_{new},j} = 1$;
    else
      $i \leftarrow$ Select a random network index from $AN_{j,k}$;
      $x^k_{i_{act},j} = 0$, $x^k_{i,j} = 1$;
    end if
  end for
end for
return A set of connections of mobiles and their services $X = \{x^1_{1,1}, \ldots, x^s_{n,m}\}$

Algorithm 5 Least Connected Algorithm

Require: List of the set of Available Networks $AN_{j,k} = \{t_1, \ldots, t_p\}, \forall j \in M, \forall k \in S$ and $p \leq n$
Require: Set of actual connection of mobiles and their services $X = \{x^1_{1,1}, \ldots, x^s_{n,m}\}$
Ensure: A (re) allocation of each used service by each connected mobile device.

$i_{new} \leftarrow -1$;
for $1 \leq j \leq m$ **do**
  for $1 \leq k \leq s$ **do**
    $i_{act} \leftarrow i \mid x^k_{i,j} = 1$;
    $i_{new} \leftarrow$ Network with less number of connections;
    if $t_{i_{new}} \in AN_{j,k}$ **then**
      $x^k_{i_{act},j} = 0$, $x^k_{i_{new},j} = 1$;
    else
      $i \leftarrow$ Select a random network index from $AN_{j,k}$;
      $x^k_{i_{act},j} = 0$, $x^k_{i,j} = 1$;
    end if
  end for
end for
return A set of connections of mobiles and their services $X = \{x^1_{1,1}, \ldots, x^s_{n,m}\}$

# Mobiles	Optimal	Without Vertical HandOver	Relative Error	Proposed Algorithm	Relative Error
10	0.852	0.231	0.729	0.717	0.158
30	0.890	0.173	0.806	0.890	0.000
50	0.991	0.179	0.819	0.767	0.226
70	0.999	0.174	0.826	0.769	0.230
100	0.955	0.176	0.816	0.882	0.076
200	1.000	0.176	0.824	0.784	0.216
300	1.000	0.172	0.828	0.841	0.159
400	1.000	0.170	0.830	0.878	0.122
500	0.945	0.169	0.821	0.939	0.006
600	0.975	0.171	0.825	0.973	0.002
700	0.991	0.172	0.826	0.990	0.001
800	0.998	0.172	0.828	0.997	0.001
900	1.000	0.173	0.827	0.999	0.001
1000	1.000	0.173	0.827	0.998	0.002

Table 6: Computed Jain's Index: Optimal, Without Vertical HandOver and using the proposed algorithm

# Mobiles	Round Robin	Relative Error	Least Connected	Relative Error
10	0.162	0.810	0.174	0.796
30	0.198	0.778	0.248	0.721
50	0.216	0.782	0.256	0.742
70	0.244	0.756	0.307	0.693
100	0.490	0.487	0.338	0.646
200	0.500	0.500	0.409	0.531
300	0.659	0.341	0.515	0.485
400	0.624	0.376	0.589	0.411
500	0.682	0.278	0.646	0.316
600	0.709	0.273	0.686	0.296
700	0.783	0.210	0.712	0.282
800	0.815	0.183	0.758	0.240
900	0.829	0.171	0.781	0.219
1000	0.860	0.140	0.737	0.203

Table 7: Computed Jain's Index: Round Robin and Least Connected Algorithms

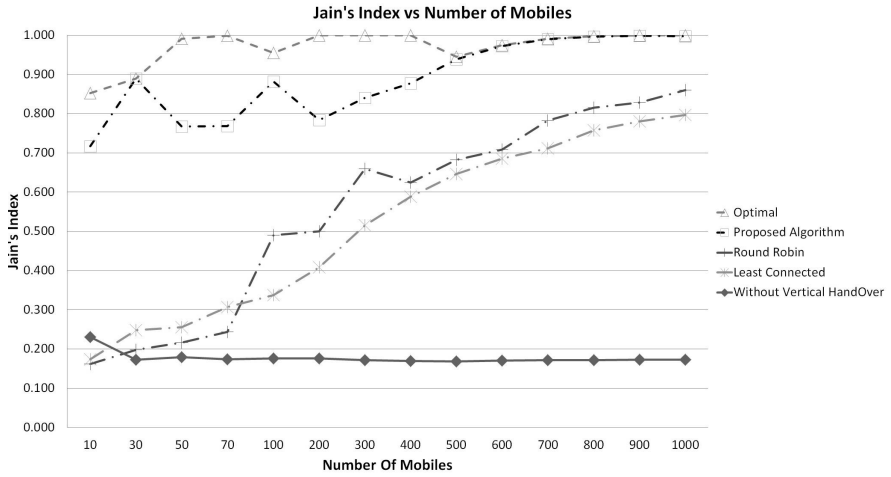


Figure 3: Effectiveness of the proposed algorithm

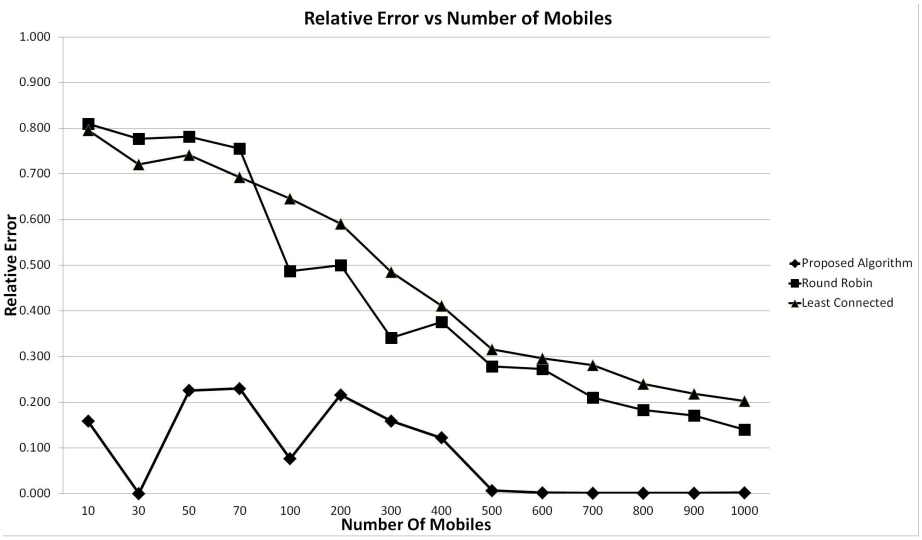


Figure 4: Relative error of the proposed algorithm

Table 8: Final network load and the computed Jain's Index for the instances of 30, 500 and 1000 mobiles

	Number of Mobiles								
	30			**500**			**1000**		
Network	**PA**	**RR**	**LC**	**PA**	**RR**	**LC**	**PA**	**RR**	**LC**
HSPA	2.14%	5.22%	11.31%	24.83%	96.08%	99.89%	48.89%	93.81%	99.11%
HSPA+	1.52%	1.90%	1.78%	23.80%	32.72%	41.53%	48.25%	76.98%	70.10%
WIMAX	1.74%	3.07%	2.53%	24.22%	39.37%	67.36%	47.89%	90.36%	99.63%
LTE	1.35%	0.70%	0.50%	23.89%	14.57%	6.01%	48.16%	29.38%	19.41%
EDGE	3.23%	68.75%	48.96%	43.75%	45.83%	94.79%	54.08%	54.17%	98.96%
WiFi N	1.36%	0.99%	0.70%	24.18%	12.46%	5.36%	48.26%	29.07%	14.02%
WiFi G	1.35%	0.74%	0.61%	24.18%	27.11%	28.72%	48.01%	55.69%	99.27%
Jain-s Index	**89.05%**	**19.84%**	**24.84%**	**93.94%**	**68.23%**	**64.61%**	**99.82%**	**86.01%**	**79.73%**

Tables 6 and 7 show the computed Jain's Index and the relative error of the solutions found by Round Robin (RR), Least Connected (LC) and the proposed (PA) algorithms. It is important to remark that for the instances between 100 and 1000 mobile devices, we relaxed the variable $x_{i,j}^k$ to accept real values between 0 and 1 in order to obtain a lower bound for those instances by nonlinear programming. As you can see in Figure 3 our proposed algorithm obtains a near optimal solutions when the number of mobiles is increased. Compared to RR and LC, we can see in Table 8 that our algorithm converge to the optimal values obtaining a fairness load balancing. The results obtained shows that the convergence of RR and LC to the optimal values of Jain's Index is due to the networks being loaded up to 100% instead of a fairness load balancing in fact.

It is appropriate to note that without vertical handover, the Jain's Index of the networks was low, and therefore the initial network load is unbalanced. Once we executed our algorithm, we observed that the resulting resource allocation had given us solutions with relative error less than 23% in the first instances. However, when the size problem increases, our algorithm presents a better convergence towards the optimal solution. Figure 4 shows how the relative error converges to zero when the number of mobiles is increased.

6 Conclusions and Future Work

In this paper, we have presented a load balancing optimization model using multihoming approach in heterogeneous wireless networks. In this approach, we have worked with different wireless access networks like HSDPA, HSPA+, Edge, WiMax, WiFi and LTE. Furthermore, in this paper we designed and implemented a Vertical Handover (VHO) Algorithm following the Always Best Connected scheme. It allows to give a solution, in a proactive way, for re-allocation of services when a new access network is available. It is also best to provide the required resources. Our proposed algorithm gives a solution to global optimality.

The resource allocation problem has been proved to be NP-Complete and we have proposed an algorithm that calculates the best load balancing solution in heterogeneous wireless networks using a multihoming approach in polynomial time $O(m^2)$.

The mathematical optimization model was computed by using the solver called BON-MIN (Basic Open-Source Nonlinear Mixed Integer Programming) and their results were compared with our proposed algorithm. Simulation results showed that without vertical handover, the initial value of the Jain's Index was very low, which means that the initial load balancing of the network was poor. However, the proposed algorithm solved the load balancing optimization model with nearly the same values as the ones given by the optimal solution obtained with the BONMIN solver.

Finally, when the number of mobiles was increased, the results obtained by our algorithm were very close to the optimal ones from the mathematical optimization model.

In further studies, we will consider researching the applicability of Evolutionary Algorithms for a multi-objective load-balancing scheme. In the technical scheme, we are going to work to implement this algorithm like a functional protocol.

Acknowledgments

The authors acknowledge to the Administrative Department of Science, Technology and Innovation (COLCIENCIAS) for the financial support to Carlos Lozano-Garzon through the 528 - 2011 National Call for Doctoral Studies in Colombia, and to the Comissionat d'Universitats i Recerca (CUR), which is part of the Departamento de innovacion, universidades i empresas (DIUE) of the Generalitat de Catalunya, the ROGER project (TEC 2012-32336) of the Spanish Government, the CSI project (SGR-1202) of the Generalitat de Catalunya, and the European Social Fund through the financial support given in the grant FI-DGR 2011 to Miguel Camelo.

Bibliography

[1] Donoso Y., Lozano-Garzon C., Camelo M., Vila P.(2014); A Multihoming Load Balancing Algorithm for a Fairness Resource Allocation in Heterogeneous Wireless Networks, International Conference on Computers, Communications & Control, Romania, Oradea, Baile Felix, May 6-10, 2014, *Abstracts of ICCCC 2014*, ISSN 1844-4334, 4: 43.

[2] Donoso Y. (2008); *Network Design for IP Convergence*, Auerbach Pubn, ISBN 978-1420067507.

[3] Global Mobile Suppliers Association, *GSM/3G Stats. Fast Facts*, Global Mobile Suppliers Association, available at `http://www.gsacom.com/news/statistics.php4`.

[4] CISCO (2013); *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017*, Cisco, San Jose, CA, USA.

[5] Kumar S., Anand S.(2011); A novel scalable software platform on android for efficient QoS on android mobile terminals based on multiple radio access technologies, *Wireless Telecommunications Symposium (WTS)*, New York City, 1-6.

[6] Ernst T., Montavont N., Wakikawa R., Ng C., Kuladinithi K. (2008); Motivations and Scenarios for Using Multiple Interfaces and Global Addresses, Draft IETF Monami6 Working Group, 2008.

[7] Lozano-Garzon C., Ortiz-Gonzalez N., Donoso Y.(2013); Mobile Network A Proactive VHD Algorithm in Heterogeneous Wireless Networks for Critical Services, *International Journal of Computers Communications & Control*, ISSN 1841-9844, 8(3): 425-431.

[8] Tversky A., Kahneman D. (1974); Judgment under Uncertainty: Heuristics and Biases, *Science*, ISSN 0036-8075, 185: 1124-1131.

[9] Donoso Y., Fabregat R.(2007); *Multi-Objective Optimization in Computer Networks Using Metaheuristics*, Auerbach Pubn, ISBN 978-0-8493-8084-6.

[10] Sousa B.M., Pentikousis K., Curado M.(2008); *Multihoming Management for Future Networks*, Mobile Network Application, ISSN 1383-469X, 16(4):505-517.

[11] Capela N., Sargento S. (2012); Optimizing Network Performance with Multihoming and Network Coding, *Globecom Workshops, 2012 IEEE*, 210-215.

[12] Ruiming Y., Yongyu C., Jia S., Dacheng Y. (2012); Traffic Split Scheme Based on Common Radio Resource Management in an Integrated LTE and HSDPA Networks, *Vehicular Technology Conference (VTC Fall), 2012 IEEE*, 1-5.

[13] Li M., Yu F., Leung, V.C.M., Randhawa T. (2004); A New Method to Support UMTS/WLAN Vertical Handover using SCT, *IEEE Wireless Communications*, 11(4):44-51.

[14] Liu B., Boukhatem N., Martins P., Bertin, P. (2010); Multihoming At Layer-2 For Inter-RAT Handover, *2010 IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, 1173-1178.

[15] Paik E.K., Heo S.Y., Kim H., Jin J.S., Lee S.C., Lee S.H. (2008); Seamless Vertical Handover Using Multihomed Mobile Access Point, *2008 IEEE Global Communications Conference*, 1-4.

[16] Folstad E.L., Helvik B.E. (2009); Managing availability in wireless inter domain access, *International Conference on Ultra Modern Telecommunications & Workshops*, 1-6.

[17] Sungwook K., Varshney, P.K. (2002); An Adaptive Bandwidth Reservation Algorithm for QoS Sensitive Multimedia Cellular Networks, *2002 IEEE 56th Vehicular Technology Conference*, 3:1475-1479.

[18] Sungwook K., Varshney, P.K. (2003); Adaptive Load Balancing with Preemption for Multimedia Cellular Networks, *2003 IEEE Wireless Communications and Networking (WCNC)*, 3: 1680-1684.

[19] Shi H., Prasad V., Onur E., Niemegeers I. (2013); Fairness in Wireless Networks:Issues, Measures and Challenges, *IEEE Communications Surveys & Tutorials*, ISSN 1553-877X, 5-24.

[20] Jain R., Chiu D.M., Hawe W.R., *A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems*, DEC Research Report TR-301, 1984.

[21] Bayrak A. E., *Optimization algorithms for resource allocation problem ff air tasking order preparation*, Master Thesis, Middle East Technical University, available at `etd.lib.metu.edu.tr/upload/12612325/index.pdf`.

[22] Mond. B, Craven B. D. (1975); Non-Linear fractional programming, *Bulletin of the Australian Mathematical Society*, 12(3) : 391-397.

[23] GAMS Development Corporation, *The General Algebraic Modeling System (GAMS)*, Retrived November 2013, from `http://www.gams.com/`.

[24] Computational Infrastructure for Operations Research, *BONMIN and BONMINH Solvers*, available at `http://www.gams.com/dd/docs/solvers/coin.pdf`.

[25] Teo Y. M., Ayani R. (2001); Comparison of Load Balancing Strategies on Cluster-based Web Servers, *Simulation*, 77(5-6): 185-195, 2001.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 9(5):570-583, October, 2014.

Trust Model in Cloud Computing Environment Based on Fuzzy Theory

L. Gu, J. Zhong, C. Wang, Z. Ni, Y. Zhang

Lichuan Gu, Chengji Wang, Youhua Zhang
School of Computer and Information, Anhui Agricultural University
No.130 ChangJiang Road, Hefei, Anhui, 230036 China
glc@ahau.edu.cn, wang_chengji@qq.com, yhzhang@ahau.edu.cn

Jinqin Zhong
University School of International Business
No.420 Linquan road, Hefei, Anhui, 230031 China.
jinqinzhong@163.com

Zhiwei Ni, Lichuan Gu
School of Management, Hefei University of Technology
No.9 Tunxi Road, Hefei, Anhui, 230009 China
gdnzw@hfut.edu.cn

Abstract: Recent years have witnessed the development of cloud computing. However, there also come some security concerns in cloud computing environment, such as emerging network attacks and intrusions, and instable cloud service provision due to flexible cloud infrastructure and resources. To this end, we research on the trusted computing in cloud computing environment. Specifically, in this paper, we propose a trust model based on virtual machines, with two considerations. First, we introduce timeliness strategy to ensure the response time and also minimize the idle time of servers. Second, we extend the linear trust chain by differentiating the trust of the platform domain and user domain. Besides, we develop a fuzzy theory based method to calculate the trust value of cloud service providers. We also conduct some experiments to evaluate our method.
Keywords: Trust model, fuzzy theory, cloud computing.

1 Introduction

Recent years have witnessed the development of cloud computing, which is an integration of parallel computing, grid computing and distributed computing [1, 2]. With massive computing and storage capability, cloud computing provides various resources to end users through trusted and reliable services. In this way, users can be relieved from trivial management routines and stay focused on the interesting business only. For example, cloud services can help to reduce the complexity of enterprise informatization process, improve the efficiency of companies' operation, and facilitate the utilization of computer resources [3].

However, although cloud computing brings us extremely convenience, there also comes some security concerns [4, 5]. The security flaws are growing ever since the complexity of system softwares increases. Also, the increasingly development of internet, as well as emerging network attacks and intrusions [6], leads to more security events. Besides, the flexibility of cloud infrastructure and resources increases the difficulty of management and brings instability. Moreover,there might be single point failure in demanding for high quality trusted service, and thus the cloud service delivery could be delayed or failed. For example, suitable authentication is required for access to bank accounts, healthrecords, intellectual property and business or politically sensitive information to reduce the security risks of cloud computing infrastructure [7].

To this end, trusted computing [8,9] is proposed by the Trusted Computing Group (TCG) [10]. With trusted computing technologies, computers can be safer and less prone to viruses and mal ware, and therefore hardware and software can consistently behave in expected manner. One way to ensure the functionality of cloud infrastructure through trusted computing is to leverage the idea of chains of trust.

In this paper, we propose a trust model and its evaluation method in cloud computing environment. Specifically, (1) the trust model in this study is built in cloud computing environment instead of traditional scenarios; (2) since time factor is significant for both QoS in requesting cloud services and also maximizing the utilization of cloud resources, we consider timeliness strategy in choosing trusted cloud services; (3) we extend the traditional linear trust chain as a tree-like structure to differentiate the trust of the platform domain and user domain; and (4) the evaluation of trusted computing is based on the fuzzy theory.

The remains of this paper are organized as follows. In Section 2 we provide some related work. Section 3 presents our proposed trust model and the fuzzy theory based evaluation method. Then experiments are conducted in Section 4. Finally, the paper is concluded in Section 5.

2 Related work

In this section, we present some related work. Generally, there are three categories: theoretical research on trusted computing (TC),architecture and implementation of TC, and TC for virtual machines environment.

The first category is theoretical efforts on trusted computing. Blaze et al. [11] first proposed the concept of trust management in 1996. Then, Josang et al. [12–14]proposed a trust model based on subjective logic, and introduced evidence space and opinion space to measure the trust relationship. Beth et al. [15] classified trust into direct trust and indirect trust, and proposed to measure the trust based on the degree of task completion. Fault tolerance capability in trusted computing within the whole life cycle of software development was also discussed [16–18]. Smith et al. [19] developed a outbound authentication model using IBM secure coprocessors. Abadi et al. [20] provided a formal description of the access control process in NGSCB system using secure logical language. Chen et al. [21]described the process of secure bootstrap in trusted computing using predicate logic.

There are also many efforts on the architecture and implementation of TC. IBM 4758 secure coprocessors [22] are one of the most earliest secure hardware. The design of IBM 4758 is to provide an isolated running environment to ensure the computing and storage capability even when something happens to the operating system or the main processors. Stanford University developed an architectural support for copy and tamper resistant software, called XOM [23]. Suh et al. [24] developed AEGIS, an architecture for a single-chip aegis processor which can be used to build computing systems secure against both physical and software attacks. Chen et al. [25] designed another secure processor Cerium. BEAR [26, 27] constructs trusted computing in commercial trusted platforms in Linux, and extends trust chain to the folder layer by checking the integrity of folders when they are first opened. Also, IBM proposed an architecture for trusted computing called IMA [28].

The last category of related work is trusted computing for virtual machines. Garfinkel et al. [29] developed a virtual machine-based platform for trusted computing, called Terra. It allows applications with a wide range of security requirements to run simultaneously on commodity hardware. Also, IBM implemented vTPM [30] to support trusted computing for multiple virtual machines, by introducing a virtual layer and extending the trust chain for virtual machines based on typical trusted platform module (TPM).

3 Proposed trust model

In typical cloud computing environment, the service model is multi-layered. That is, the cloud service provider (CSP) not only provides services to end users, but also to the upper layer CSPs, which forms a trust chain of service providers and consumers. As shown in Figure 1, the arrow denotes from providers to consumers. For example, the service of end users might be provided by a SaaS CSP directly, or first PaaS CSP then SaaS CSP, or from IaaS CSP downstream to end users.

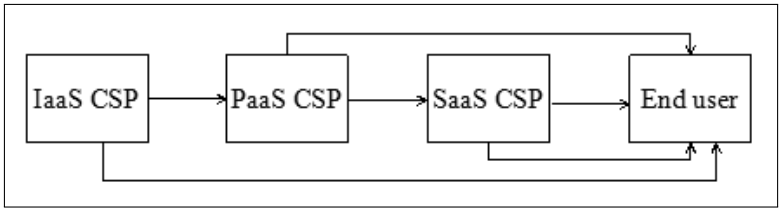


Figure 1: Trust transitivity service providers and consumers

There are two kinds of trust in cloud computing environment: direct and indirect trust. Direct trust means the impression of consumer users on the service quality of CSP, while indirect trust denotes the aggregated impression of all other previous consumers who have used the service or other CSPs who have connections with current CSP.

Suppose user u_a wants to use the cloud service from CSP X .The objective is to calculate the trust value of X and determine if X is trusted. Denote $\mathrm{Imp}(a, b)$ as the impression of b on a. Therefore, the trust value of X for u_a can be calculated as:

$$\mathrm{Imp}(X, u_a) = C_1 \sum_{u_i \in U} \alpha_i \mathrm{Imp}(X, u_i) + C_2 \sum_{Y} \beta_i \mathrm{Imp}(X, Y) \tag{1}$$

where $\mathrm{Imp}(X, u_i)$ is the impression value of u_i on CSP X , $\mathrm{Imp}(X, Y)$ is the impression of other CSPs on X, and C_1, C_2, α_i, β_i are coefficients.

3.1 Timeliness strategy

In a cloud computing environment, host data nodes provide trusted services with high credible and stable resources, where timeliness is a significant indicator. Generally, the process of trusted cloud services is as follows. First, if there exist idle host data nodes, then evaluate the timeliness of the node. Second, calculate the trust value of data node based on the timeliness and trust model. If the trust value is satisfactory, then the node is assigned for cloud services; otherwise, repeat the process for the next service request.

We assume that the trust value is related to time, and the more recent the evaluation is, the more contribution it has to the current trust calculation. Let $\mathrm{Imp}(X, u_i)_L$ be the trust value at time L Therefore, the trust value of X for u_a can be represented by rewriting Equation (1) through introducing timeliness strategy:

$$\mathrm{Trust}(X, u_a) = \mathrm{Imp}(X, u_a)_L - \Delta\mathrm{Imp} \cdot F(t - L) \tag{2}$$

where $\mathrm{Imp}(X, u_a)_L$ is the current impression value at time L , and $\Delta\mathrm{Imp} \cdot F(t - L)$ indicates the affect of previous trust value.

Suppose there are k historical impression values before L, and $\mathrm{Imp}(X, u_a)_L$ is defined as:

$$\mathrm{Imp}(X, u_a)_L = \sum_{i=1}^{k} \mathrm{Imp}(X, u_a)_i w_i \tag{3}$$

where $\sum_{i=1}^{k} w_i = 1$.

Since the history impression evaluation happens randomly, we equally split the history time period into m segments in order to differentiate the importance of each historical evaluation. Historical impressions within the same time window W are assigned an identical weight ω_W . Therefore, inspired by [31], weight w_i can be calculated as:

$$w_i = \sum_{j=1}^{n} \frac{\sqrt[m-1]{\omega_{W_1}(m-j)\omega_{W_m}(j-1)}}{n} \tag{4}$$

where n is the number of values within i-th historical impression. Substitute Equation (4) into (3), we get

$$\mathrm{Imp}(X, u_a)_L = \sum_{i=1}^{k} \mathrm{Imp}(X, u_a)_i \sum_{j=1}^{n} \frac{\sqrt[m-1]{\omega_{W_1}(m-j)\omega_{W_m}(j-1)}}{n} \tag{5}$$

Suppose the experience distribution of ΔImp follows Gumbel distribution [32]. Therefore the estimate of ΔImp is:

$$\hat{G}(\mathrm{Imp}_i) = \exp\left\{ -\exp\left\{ -\frac{\mathrm{Imp}_i - \hat{u}}{\hat{\delta}} \right\} \right\} \tag{6}$$

where $\hat{u}$, $\hat{\delta}$ is the maximum likelihood estimates for Gumbel distribution.

Time based function $F(t-L)$ follows exponential distribution:

$$F(t-L) = \int_{L}^{+\infty} (t-L) f(t) dt = \frac{\exp(-\lambda L)}{\lambda} \tag{7}$$

3.2 Tree structured trust chain

In this section, we consider the trust between a specific CSP and users. The objective is to measure the trust value from the CSP hardware to user defined software.

Trust relationship in cloud environment is more complicated than the traditional scenarios. For example, the services are typically running in virtual machines with larger management and user domain, which is hardly measured by the traditional linear structure. Moreover, for some business requirements, services are assembled across multiple user domains, which also increase the difficulty of measuring the trust. Note that by management domain, we mean the set of objects, typically refers to the component in traditional trust chain; by user domain, we mean that different users that request for the service.

Typically, cloud users don't have controls over the hardware devices, and thus the safety is only ensured by service-level agreement (SLA). However, users wish to somehow control the virtual computing resources, i.e., virtual machines (VM), so that user-defined security strategy can be passed over VM and therefore ensure the safety of cloud resources.

One popular solution is to use trusted platform module (TPM), which is a specialized chip that can securely store information, such as passwords and encryption keys, with independent execution CPU unit. The typical architecture of TPM is shown in Figure 2. TPM provides trusted computing by ensuring security to operating systems and TCG Software Stack (TSS). TSS is a support software of TPM, and its architecture is shown in Figure 3.

In cloud computing environment, trusted computing is typically implemented by virtualizing the Trusted Platform Module (vTPM). However, vTPM is only a virtual instance of physical TPM in user domain, which calls the physical TPM resource to provide TPM service, as shown

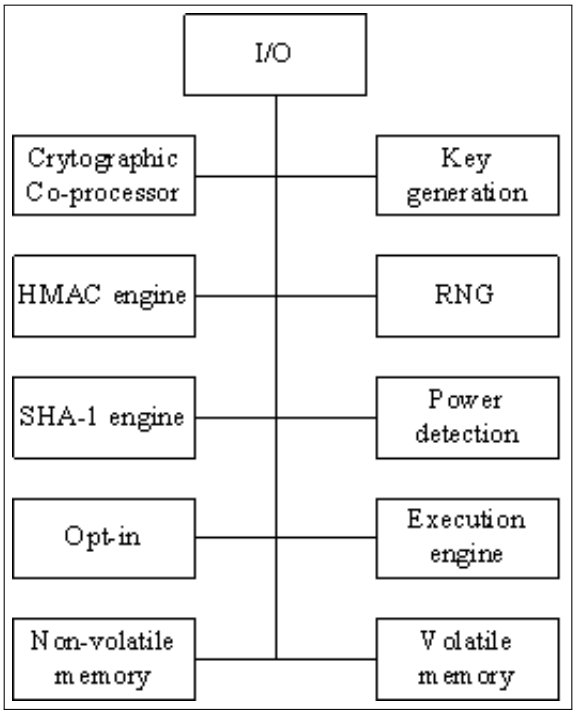


Figure 2: TPM component architecture

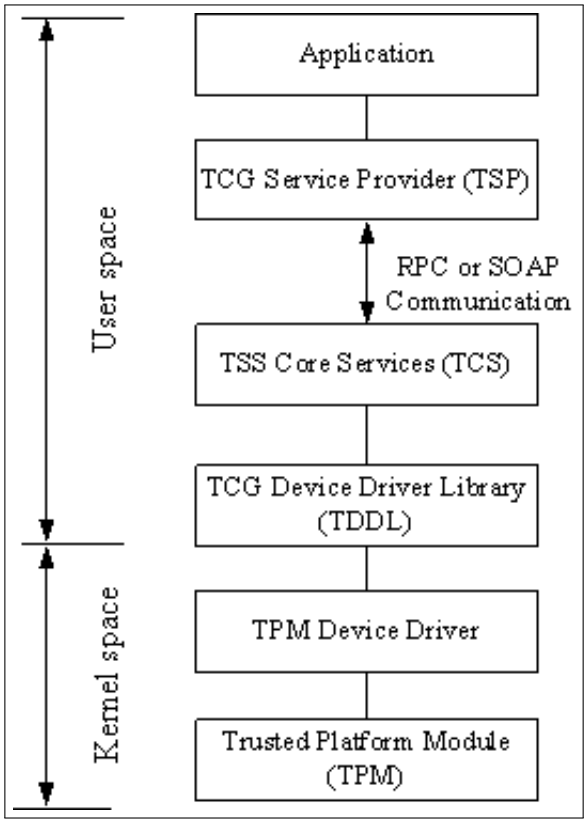


Figure 3: Architecture of TSS

in Figure 4. There are some limitations of this simple structure. First, the whole structure is dependent on limited physical TPM, which is not scalable for large scale virtual machines in cloud platform. Second, evaluation of each individual vTPM is sequential due to the limited physical resource, and therefore it is not dynamic and elastic. Last, vTPM relies on physical TPM, that is, the trusted capability is ensured by the cloud infrastructure only. Accordingly, users cannot specify their personalized security strategies on demand. Moreover, we observe that existing trust models, where the evaluation is all performed by one single node, are inappropriate especially in cloud computing environment [33].

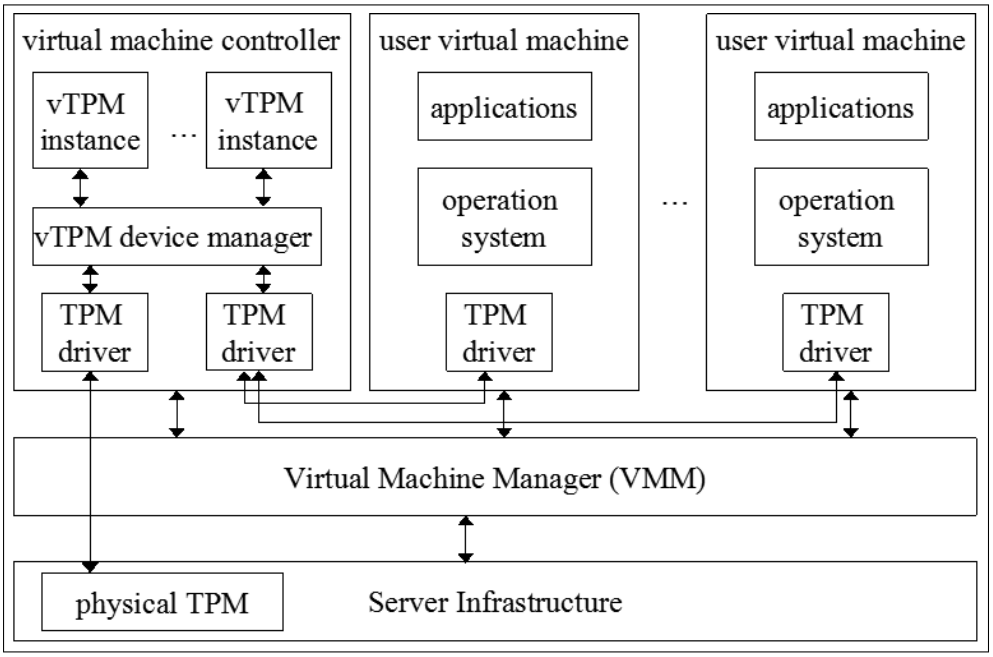


Figure 4: Illustration of user virtual machine

To this end, in this section, we propose a tree structured trust model by distributing the evaluation work to multiple nodes. Specifically, we combine the trust evaluation from physical platform and user domain. As shown in Figure 5, there are two stages in the evaluating of trust value. First, there is a chain of trust in physical TPM, i.e., CRTM'BIOS'GRUB'VMM. The integrity and security of the system is ensured by the isolation mechanism of physical TPM and cloud infrastructure. Second, TPM controller creates TPM for each user, i.e., uTPM, which is responsible for the evaluation and security of software components in the user domain. Each uTPM is created for a user virtual machine, and holds the results of integrity evaluation and reports to users.

There are two kinds of trust transitivity in this model. (1) Pass the trust from physical TPM to TPM controller, from hardware to user virtual machine. In this way, we can combine the trust of platform and the trust of user virtual machine together, to provide a complete trust chain. (2) Pass the trust from user to virtual machine. By virtualizing TPM, an independent vTPM is created for each user virtual machine, i.e., uTPM. In Figure 5, each node in the upper physical trust chain is responsible for evaluating trust, and the evaluation is performed for each user in parallel in the lower user domain trust chain.

Theorem 1. *The tree structured trust model in Figure is trusted.*

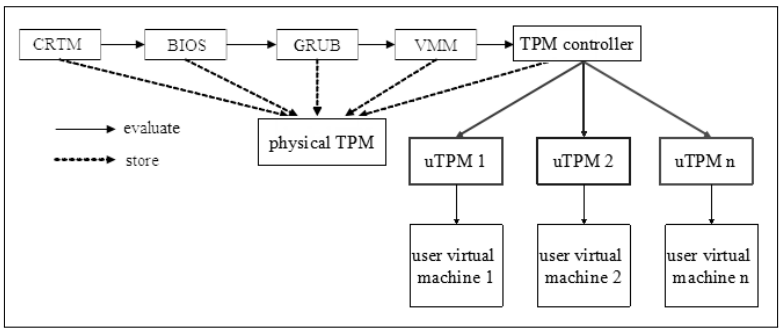


Figure 5: Illustration of tree structured trust model

Proof: For ease of description, we simplify the tree structured trust model in Figure 6. Denote the trust value of node A for B as $T(A,B)$. Inferred by the trust transitivity principle, we have $T(A,B) = \min\{T(A,B), T(B,C), \ldots, T(N-1,N)\}$. As proved in [24], the left part chain $A \to B \to C \to \cdots \to N$ is trusted.

Now we consider the right part chain. Each node U_i represents a trust chain of each user virtual machine, and therefore U_i itself is trusted. There are multiple direct paths, i.e., $T_1(N,U_1), T_2(N,U_2), \ldots, T_N(N,U_N)$. The final trust value $T(N,U)$ is no less than

$$\max\left\{T_1(N,U_1), T_2(N,U_2), \ldots, T_N(N,U_N)\right\}.$$

Therefore the right part chain is trusted. □

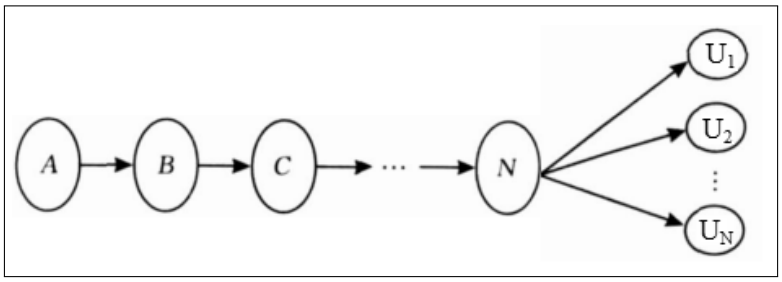


Figure 6: Illustration of tree structured trust model

3.3 Fuzzy theory based trust model

As used in many Suppose the confidence level for trust is $U = \{U_1, U_2, U_3, U_4\} = \{$'not trusted', 'somehow trusted', 'normal trusted', 'complete trusted'$\}$, and the confidence vector $V = \{v_1, v_2, v_3, v_4\}$, where $v_i (i = 1, 2, 3, 4)$ denotes the degree of membership of each level U_i.

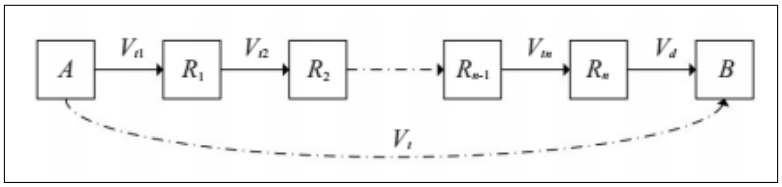


Figure 7: Illustration of trust transitivity

The trust between initial entity and target entity is evaluated through the stepwise evaluation and the transitivity between them. As shown in Figure 7, where A is observed entity, B is evaluation entity, and $R_1, R_2, \ldots, R_n$ are the intermediate entities. The corresponding trusts between nodes are $V_{t1}, V_{t2}, \ldots, V_{tn}, V_d$, and the trust vector from A to B is V_t . The integrity of direct measurement is mainly affected by the measuring capability of the current evaluation node.

Now we consider the timeliness factor. Let $DT(R_i, Rj, t)$ be the direct trust of R_j in R_i at time t , and it can be calculated as:

$$DT(R_i, R_j, t) = \frac{S_{R_j R_i}}{S_{R_j R_i} + F_{R_j R_i}} \delta(t, t_0) \tag{8}$$

where $S_{R_j R_i}$ denotes the number of successful historical integrity evaluation of R_j on R_i, $F_{R_j R_i}$ is the number of failure evaluation, and t, t_0 denote the current and first evaluation time respectively. $\delta(t, t_0)$ is the time decay function, defined as:

$$\delta(t, t_0) = 1 - \frac{t - t_0}{t} \xi \tag{9}$$

where $\xi \in [0, 1]$ is the adjustment factor of the decay.

The value of direct trust $DT(R_i, R_j, t)$ can be transformed into a fuzzy vector $V(i, j) = (v_1, v_2, v_3, v_4)$, where

$$v_1 = \begin{cases} 1 - 2\left(\frac{DT}{0.5}\right)^2, & 0 \le DT \le 0.25; \\ 2\left(\frac{0.5 - DT}{0.5}\right)^2, & 0.25 < DT < 0.5; \\ 0, & 0.5 < DT < 1. \end{cases} \tag{10}$$

$$v_2 = \begin{cases} 0, & 0 \le DT \le 0.25; \\ 1 - 2\left(\frac{DT - 0.25}{0.5}\right)^2, & 0.25 < DT \le 0.25; \\ 2\left(\frac{0.75 - DT}{0.5}\right)^2, & 0.5 < DT \le 0.75; \\ 0, & 0.75 < DT \le 1. \end{cases} \tag{11}$$

$$v_3 = \begin{cases} 0, & 0 \le DT \le 0.25; \\ 2\left(\frac{DT - 0.25}{0.5}\right)^2 & 0.25 < DT \le 0.5; \\ 1 - 2\left(\frac{0.75 - DT}{0.5}\right)^2 & 0.5 < DT \le 0.75; \\ 0 & 0.75 < DT \le 1. \end{cases} \tag{12}$$

$$v_4 = \begin{cases} 0, & 0 \le DT \le 0.5; \\ 2\left(\frac{DT - 0.5}{0.5}\right)^2 & 0.5 < DT \le 0.75; \\ 1 - 2\left(\frac{DT - 1}{0.5}\right)^2 & 0.75 < DT \le 1. \end{cases} \tag{13}$$

Suppose there are n evaluation metrics $Z = \{z_1, z_2, \ldots, z_n\}$, we get the evaluation matrix for entity R_i :

$$V_i = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ v_{3,1} & v_{3,2} & \cdots & v_{3,n} \\ v_{4,1} & v_{4,2} & \cdots & v_{4,n} \end{bmatrix} \tag{14}$$

Therefore, the indirect trust of R_i is calculated as:

$$IT_i = W_i \cdot V_i \tag{15}$$

where $W_i = (w_1, w_2, w_3, w_4)$ is the weights of each metric, and $w_i \in [0, 1]$, $\sum_{i=1}^{4} w_i = 1$. We employ a fuzzy reasoning method based on similarity. Suppose there exist a rule:

$$R : \text{If } A \text{ Then } B, \lambda, W \tag{16}$$

where A is the antecedent component, B is the consequent component of the rule, and λ is a threshold which decides the rule to be executed or not.

The similarity between two entities is calculated as:

$$S(A\prime_i, A_i) = \begin{cases} \frac{M(A\prime_i \cap A_i)}{M(A\prime_i) \vee M(A_i)}, & \text{if } A\prime_i \subseteq A_i \text{ or } A\prime_i \supseteq A_i \\ \frac{A\prime_i \cap A_i}{M(A\prime_i)}, & \text{else.} \end{cases} \tag{17}$$

where $A\prime_i$, A_i are fuzzy set, $M(A_i) = \sum_{x \in X_i} \mu_{A_i}(x)$, X_i is the mathematical domain of discourse for $A\prime_i$, A_i.

As shown in Figure, in our case, the antecedent components are $V_{t1}, V_{t2}, \ldots, V_{tn}$, and the consequent component is V_t. That is, the rule is:

$$\begin{aligned} R : &\text{ If the measurement of } A \text{ for } R_1 \text{ is } V_{t1} \\ &\text{AND the measurement of } R_1 \text{ for } R_2 \text{ is } V_{t2} \\ &\text{AND } \cdots \\ &\text{AND the measurement of } R_n \text{ for } B \text{ is } V_d \\ &\text{Then the measurement of } A \text{ for } B \text{ is } V_t \end{aligned} \tag{18}$$

Given the observed values are $V\prime_{t1}, V\prime_{t2}, \ldots, V\prime_{tn}$, the goal is to calculate the observed $V\prime_t$. The process of inference is as follows.

Step 1: Calculate the similarity between observed $S(V\prime_{ti}, V_{ti}), i = 1, 2, \cdots, n$;

Step 2: If $S(V\prime_{ti}, V_{ti}) > \lambda$, calculate the overall similarity:

$$S_W(V\prime_{ti}, V_{ti}) = \sum_{i=1}^{n} S(V\prime_{ti}, V_{ti})^* \frac{w_i}{\sum_{j=1}^{n} w_j} \tag{19}$$

Step 3: Compute the inference results:

$$\theta_1 = \sum_{i=1}^{k_1} \frac{S(V\prime_{ti}, V_{ti})^* w_i}{\sum_{i=1}^{k_1} w_j} \tag{20}$$

where $j \in \{i : M(V\prime_{ti}) \geq W(V_{ti})\}$, and

$$\theta_2 = \sum_{i=1}^{k_2} \frac{S(V\prime_{ti}, V_{ti})^* w_i}{\sum_{i=1}^{k_2} w_j} \tag{21}$$

where $j \in \{i : M(V\prime_{ti}) < W(V_{ti})\}$.
If $\theta \neq 0$ and $\theta_2 \neq 0$, then

$$V\prime_t = \begin{cases} \frac{V_t * \theta_1}{\theta_2}, & \theta_1 \leq \theta_2; \\ \min\left\{1, \frac{V_t * \theta_1}{\theta_2}\right\}, & \text{otherwise.} \end{cases} \tag{22}$$

If $\theta_1 = 0$ or $\theta_2 = 0$, then

$$V_t = \begin{cases} V_{t^*} S_W, & \theta_1 \leq \theta_2; \\ \min\left\{1, \frac{V_t}{S_W}\right\}, & \text{otherwise.} \end{cases} \tag{23}$$

To sum up, in this section we presented the proposed trust model, which works as follows. When user a wants to use some service, for each qualified candidate CSP X, he/she first inquires the information from other users and CSPs who have interactions with X. Then, for each X, a tree structured trust chain is constructed, where the upper part is the trust chain for physical TPM of X, and the lower part is built for each possible user. After that, an evaluation method based on fuzzy theory is performed to calculate the trust value of a for X. Once all the trust value is learned, user a can determine if a CSP X is trusted or not.

4 Experiment

In this section, we evaluate the efficiency of our proposed tree structured trust model. The configuration of PC is as follows. Intel Core i5 2.8 GHz CPU with four cores, 4 GB memory, 500 G hard disk. We use Xen virtualization platform for virtualization implementation, CloudSim [34] for cloud computing platform simulation, and Matlab for fuzzy system implementation.

4.1 Evaluating trusted cloud service selection

Figure 8 shows the trust value of four types of CSPs with different number of transactions. We have the following observations. First, for complete trusted and normal trusted CSPs, the trust value is near linearly growing since they are offering real trusted services. Second, for somehow trusted CSPs, the trust value is unstable. Third, the value of not trusted CSPs decreases quickly, and they would not be selected as the cloud service provider.

4.2 Evaluating timeliness

We compare our method with traditional Dynamic Level Scheduling (DLS) algorithm [35] for scheduling cloud services with the consideration of timeliness factor. The results are reflected as the average of 100 executions.

Figure 9 shows the ratio of successful execution with different numbers of tasks. We can observe that the successful execution tends to increase when the number of tasks is growing. However, for DLS algorithm, the maximum success ratio is around 0.65. The reason is that it does not consider the node failure at specific time. By contrast, our method performs better because we consider the timeliness factor in all historical evaluations.

Figure 10 shows the average schedule length with different numbers of tasks. The average schedule length grows when there are more tasks. Also, the average schedule length of DLS is smaller than that of our method. Therefore, we can see that our method achieves larger success ratio by sacrificing the scheduling time. Although the time cost increases, our method can help to select more successful trusted cloud services.

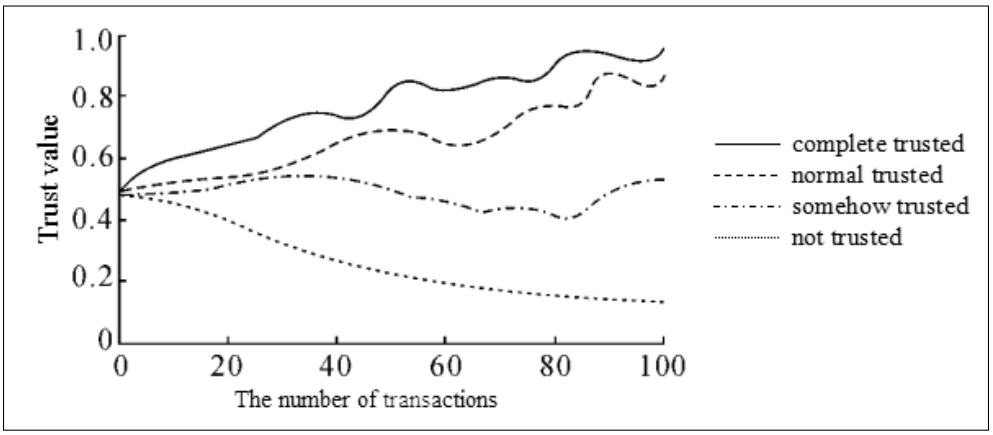


Figure 8: Trust value of four types with different number of transactions

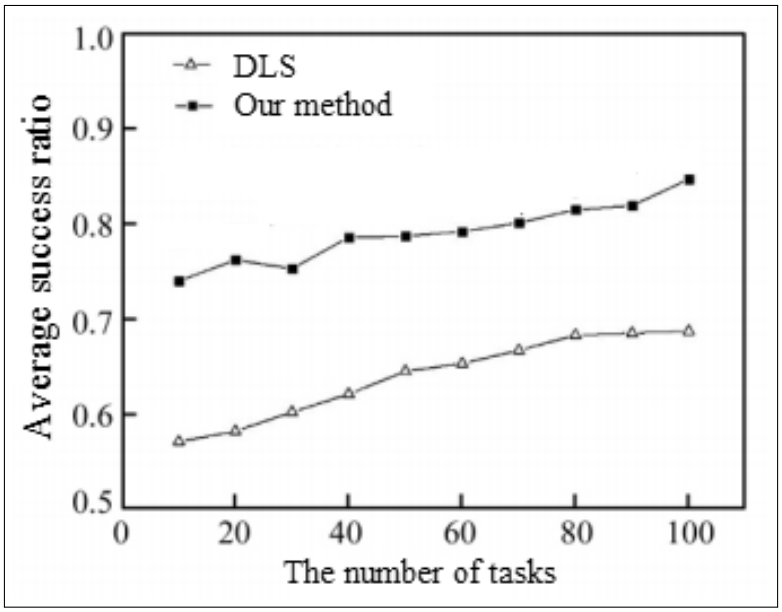


Figure 9: The ratio of successful execution with different numbers of tasks

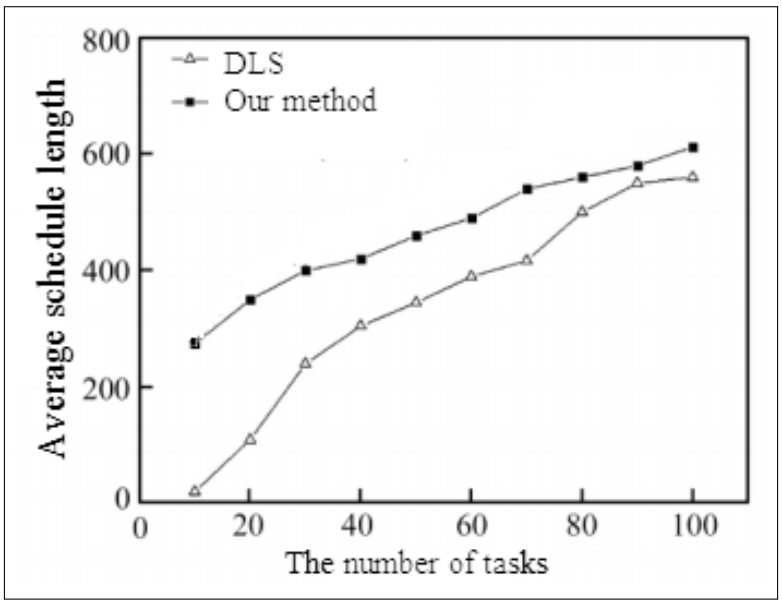


Figure 10: Average schedule length with different numbers of tasks

5 Conclusion

In this paper, we proposed a trust model in cloud computing environment. Specifically, the trust model is designed for virtual machines with the consideration of timeliness factor. Moreover, we employ a fuzzy theory based method to calculate the trust value of specific CSP.

In our experiments, we exhibit the trust value of four pre-defined confidence levels, and also evaluate the efficiency of the timeliness consideration. We find that our method can improve the successful response of selecting cloud services at the expense of average schedule length.

However, in future works, we might want to explore a more optimal balance between efficiency and effectiveness.

Acknowledgements

This work was supported by the National Natural Science Foundation of china (Grant No.313715
33, Grant No. 71271071), National Science and Technology Support Program (Grant No. 2013BAJ10B12), the Key Technologies R & D Program of AnHui Province (Grant No.1301032169) and by the Natural Science Foundation of AnHui Province (Grant No.1308085MF89).

Bibliography

[1] Armbrust, Michael, et al. (2010); A view of cloud computing, *Communications of the ACM*, 53(4): 50–58.

[2] Mell P., Grance T. (2011); The NIST definition of cloud computing, `http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf`, 1-7.

[3] Lin C., Pervan G. (2001); A review of IS/IT investment evaluation and benefits management issues, problems and processes, in *Information technology evaluation methods and management*, ISBN:1-878289-90-X, 2-24.

[4] Brodkin J. (2008); Gartner: Seven cloud-computing security risks. *Infoworld* (2008): 1–3.

[5] Zissis D., Lekkas D. (2012); Addressing cloud computing security issues. *Future Generation Computer Systems*, 28(3): 583–592.

[6] Lonea A.M., Popescu D.E., Tianfield H.(2012); Detecting DDoS Attacks in Cloud Computing Environment, *International Journal of Computers Communications & Control*, 8(1): 70–78.

[7] Popescu D.E, , Lonea A.M. (2013); An Hybrid Text-Image Based Authentication for Cloud Services, *International Journal of Computers Communications & Control*, 8(2): 263–274.

[8] Pearson S., Balacheff B., eds. (2003); *Trusted computing platforms: TCPA technology in context*, Prentice Hall Professional.

[9] Mitchell C. ed.(2005), *Trusted computing*, Institution of Electrical Engineers.

[10] Sumrall N., Novoa M. (2003); Trusted Computing Group (TCG) and the TPM 1.2 Specification. *Intel Developer Forum.* Vol. 32.

[11] Blaze M., Feigenbaum J., Lacy J. (1996); Decentralized trust management. *Security and Privacy, 1996 IEEE Symposium on*, 164-173.

[12] Josang A. (2001); A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3): 279–311.

[13] Knapskog, S. J.(1998); A metric for trusted systems, *Proc. of the 21st National Security Conference*, Available at `http://folk.uio.no/josang/papers/JK1998-NSC.pdf`, 1-14.

[14] Josang A. (1999); Trust-based decision making for electronic transactions, *Proc. of the Fourth Nordic Workshop on Secure Computer Systems*, 1-21.

[15] Beth T., Borcherding M., Klein B. (1994); *Valuation of trust in open networks*, Springer Berlin Heidelberg.

[16] Meyer J. F. (1980); On evaluating the performability of degradable computing systems. *Computers, IEEE Transactions on*, 100(8): 720–731.

[17] Isermann R. (1984); Process fault detection based on modeling and estimation methods, a survey. *Automatica*, 20(4): 387–404.

[18] Arlat J. et al.(1993); Fault injection and dependability evaluation of fault-tolerant systems, *Computers, IEEE Transactions on*, 42(8): 913–923.

[19] Smith S. W. (2002); Outbound authentication for programmable secure coprocessors. *Computer Security, ESORICS, 2002. Springer Berlin Heidelberg*, 72–89.

[20] Abadi M., Wobber T. (2004); A logical account of NGSCB. Formal Techniques for Networked and Distributed Systems, CFORTE 2004. Springer Berlin Heidelberg, 2004. 1–12.

[21] Chen S., Wen Y., Zhao H. (2007); Formal analysis of secure bootstrap in trusted computing, *Autonomic and Trusted Computing*, Springer Berlin Heidelberg, 352–360.

[22] Dyer J. G., et al. (2001); Building the IBM 4758 secure coprocessor. *Computer*, 34(10): 57–66.

[23] Lie, David, et al.(2000); Architectural support for copy and tamper resistant software, *ACM SIGPLAN Notices* , 35(11): 168–177.

[24] Suh G. E. et al. (2003); AEGIS: architecture for tamper-evident and tamper-resistant processing. *Proc. of the 17th annual international conference on Supercomputing. ACM*, 1-18.

[25] Chen B., Morris R. (2003); Certifying Program Execution with Secure Processors, *HotOS*, Available at `http://pdos.csail.mit.edu/papers/cerium:hotos03.pdf`, 1-6.

[26] MacDonald R. et al. (2003); Bear: An open-source virtual secure coprocessor based on TCPA. *Computer Science Technical Report TR2003-471*, Dartmouth College.

[27] Marchesini J. et al.(2003); Experimenting with TCPA/TCG hardware, or: How I learned to stop worrying and love the bear. *Computer Science Technical Report TR2003-476*, Dartmouth College .

[28] Sailer R. et al. (2004); Design and Implementation of a TCG-based Integrity Measurement Architecture. *USENIX Security Symposium*, 13:223-238.

[29] Garfinkel T. et al.(2003); Terra: A virtual machine-based platform for trusted computing, *ACM SIGOPS Operating Systems Review*. 37(5):193-206.

[30] Berger S. et al. (2006); vTPM: virtualizing the trusted platform module, *Proc. 15th Conf. on USENIX Security Symposium*, 305-320.

[31] Fullér R., Majlender P. (2001); An analytic approach for obtaining maximal entropy OWA operator weights, *Fuzzy Sets and Systems*, 124(1): 53–57.

[32] Saure D. et al. (2010); Time-of-use pricing policies for offering cloud computing as a service, *Service Operations and Logistics and Informatics (SOLI), 2010 IEEE International Conference on*, 300-305.

[33] Bo Z. et al.(2010); The system architecture and security structure of trusted PDA, *Chinese Journal of Computers*, 33(1): 82–92.

[34] Calheiros R. N. et al.(2011); CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and Experience*, 41(1): 23–50.

[35] Wang W. et al. (2012); Dynamic trust evaluation and scheduling framework for cloud computing. *Security and Communication Networks*, 5(3): 311–318.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 9(5):584-592, October, 2014.

Choice of Countermeasures in Project Risk Management Using Fuzzy Modelling

D. Kuchta, D. Skorupka

Dorota Kuchta*
1. Wroclaw University of Technology
Poland, 50-370 Wroclaw, Wybrzeze Wyspianskiego 27,
dorota.kuchta@pwr.wroc.pl
2. Tadeusz Kosciuszko Military Academy of Land Forces
Poland, 51-150 Wroclaw, ul. Czajkowskiego 109,
*Corresponding author

Dariusz Skorupka
Tadeusz Kosciuszko Military Academy of Land Forces
Poland, Wroclaw

Abstract: The paper proposes a new method for project risk management. It is proposed how, after risk identification, the countermeasures for risk mitigation and elimination can be selected, taking into account the cost and effort linked to them as well as the weights assigned by the decision maker to risk attributes, such as probability or consequences, and the values of those attributes. The risk attributes and weights, as well as the maximal total risk and the maximal total effort of risk mitigation accepted by the decision maker for the project are expressed as fuzzy numbers, which in turn constitute models for linguistic expressions.
Keywords: project risk, risk mitigation, risk transfer, risk elimination, fuzzy logic-based optimisation.

1 Introduction

Project risk management is a very important subject and there is a lot of literature connected to it. There are also a lot of risk definitions (it is assumed here, after [1], that risk is defined as an event which may happen and if it does, it will have negative consequences on at least one of project success determinants[1]) and risk management methods or techniques (e.g. [1]). Most of them can be summarized as follows:

I. Risk identification;

II. Risk evaluation - the risks identified in point I. are evaluated on the basis of their attributes (like probability and consequences);

III. Risk elimination, transfer or mitigation - taking some steps or countermeasures which eliminate some risks or decrease the evaluation of others;

IV. Risk control during the project execution;

V. Lesson learned recording.

The problem to which this paper wants to contribute is point III. of the above procedure. There are no formal methods in the literature, apart from one presented in [10], helping the decision maker in this step. Normally only informal procedures are proposed (e.g. [6]). And in [10] the authors assume among others that the probability of the risks cannot be explicitly

[1]Project success determinants are discussed later in the paper.

influenced by the countermeasures, but only the expected value of monetary losses, which in our opinion is wrong - in reality usually each risk attribute (e.g. probability, consequences, etc.), can be individually influenced by individual countermeasures. The authors of [10] assume also that the goal is the maximization of the monetary expected value of risk reduced by the countermeasures, whereas we assume that the goal should be the minimization of countermeasures total cost needed to attain a fixed total risk level of the project, which we think to be more realistic. Also, in [10] crisp values are required for risk attributes, ever for quality, whereas we assume linguistic expressions (fuzzy numbers) for risk attributes, like it is done e.g. in [4] and [9], as we think crisp values would usually be difficult to obtain from the decision maker. Also, basing ourselves on [9], it is assumed here that different risk attributes may have various weights in the eyes in the decision maker, which in [10] is ignored. For one decision maker mostly the consequences may count (this is the case for projects where consequences are injuries or deaths of human beings), for some decision makers mostly the probabilities (it is so for many public projects, where each problem, even a relatively small one, is noticed by the media who make a big deal of it). What is more, nowhere in the literature the notion of risk levels (e.g. project level, closest environment level, further environment level, national level, international level etc., introduced in [8]) and the effort linked to an effective application of countermeasures on each of the different levels are combined with the choice of risk countermeasures, and this is done in the present paper.

2 Basic notions and notation

As it was mentioned above, project risk is an event which might happen and if it happens, it will have negative consequences on one of the success determinants of the project. There are several determinants of project success (for a review see e.g. [2]), like time, quality, cost, customer satisfaction, etc. Let us denote those determinants as D_l, $l = 1, \ldots, L$.

The risks may concern several levels ([8]), e.g. the project level, the closer environment level (for construction projects this would be the construction market level), the further environment level (for construction projects this would be the national level), etc. The various levels are denoted as L_t, $t = 1, \ldots, T$. The higher the level index, the harder accessible and the less possible to influence the level is, thus the less manageable the risks linked to this level are.

It is assumed that (point I. of the project risk management procedure given above) P_t^l risks for each level L_t, $t = 1, \ldots, T$ and for each determinant D_l, $l = 1, \ldots, L$ have been identified (the risks will be denoted as $R_k^{t,l}$, $k = 1, \ldots, P_t^l$). Each risk $R_k^{t,l}$, if it happens, it concerns level L_t (thus, if one wanted to mitigate or eliminate it, one would have to act on the respective level) and will influence determinant D_l, $l = 1, \ldots, L$ of project success. It is also assumed that each risk is linked to exactly one level and to exactly one determinant, but this assumption would not be difficult to be given up.

Each risk $R_k^{t,l}$ will have several attributes. Usually two or at the most three risk attributes are considered: the probability of occurrence, the consequences of the occurrence (for the project success determinant linked to the specific risk, i.e. risks may have consequences for time, cost, quality, customer satisfaction etc.) and sometimes the difficulty in an early detection or in early forecasting of the risk occurrence (e.g. [6]). Sometimes additional attributes may be important, like the probability of the occurrence of consequences once the risk has happened, the degree of influence we have on the occurrence of the risk etc. The latter attribute is especially important in case we have to consider several levels on which the risks may occur: the further the level, the less effective our countermeasures may be.

Let us denote the number of attributes as A, and the attributes themselves as v_a, $a = 1, \ldots, A$. The attributes are functions of risks, thus the value of the a-th attribute of risk $R_k^{t,l}$ will be

$v_a(R_k^{t,l})$, $a = 1, \ldots, A\,, l = 1, \ldots, L$, $t = 1, \ldots, T$, $k = 1, \ldots, P_t^l$. $v_a(R_k^{t,l})$ will be all measured in the same scale, their values will be given by experts.

As mentioned above, various risk attributes may have various weights in the eyes of the decision maker. Each attribute $v_a(R_k^{t,l})$ will be thus assigned a weight $w_a(R_k^{t,l})$, also according to a given scale.

Finally, it is possible to use the defined parameters to evaluate each risk (point II. of the project risk management procedure). The evaluation $REV(R_k^{t,l})$ of each risk may be defined using one of the formulae proposed e.g. in [6][2], e.g. the following one:

$$REV(R_k^{t,l}) = \frac{\sum_{a=1}^{A} w_a(R_k^{t,l}) \cdot v_a(R_k^{t,l})}{\sum_{a=1}^{A} w_a(R_k^{t,l})} \tag{1}$$

As a further step (point III. of the project risk management procedure), it is proposed to try to identify certain countermeasures to eliminate or mitigate the risks. It is assumed that each measure may concern only one level, one project success determinant and one risk attribute. This assumption seems natural: one countermeasure can e.g. be effective only on the national level and affect only the time, i.e. a possible delay being the consequence of a certain risk, another countermeasure may concern only the project level and affect the cost and the probability of an event which would increase cost. Let us introduce also a simplifying assumption that each attribute of each risk can be affected by only one measure. Thus, for each $l = 1, \ldots, L$, $t = 1, \ldots, T$ and $a = 1, \ldots, A$ let us define a set of measures $(M_h^{t,l,a}, h = 1, \ldots, H_t^{l,a})$. Then, for each attribute $v_a(R_k^{t,l})$, $a = 1, \ldots, A$ of each risk $R_k^{t,l}$ $(k = 1, \ldots, P_t^l)$ it is identified which measure among the measures $\left\{M_h^{t,l,a},\ h = 1, \ldots, H_t^{l,a}\right\}$ can be applied, and the index of this measure will be denoted as $M(v_a(R_k^{t,l}))$. The value of each attribute of the risk $R_k^{t,l}$ after the application of the measure indexed by $M(v_a(R_k^{t,l}))$ will be equal to $m(v_a(R_k^{t,l}))$. Of course it is true $m(v_a(R_k^{t,l})) < v_a(R_k^{t,l})$. For some risks and attributes no countermeasure may exist - then it is set everywhere $m(v_a(R_k^{t,l})) = v_a(R_k^{t,l})$.

The evaluation of each risk after the application of the measures will be

$$REV(R_k^{t,l}) = \frac{\sum_{a=1}^{A} w_a(R_k^{t,l}) \cdot m(v_a(R_k^{t,l}))}{\sum_{a=1}^{A} w_a(R_k^{t,l})} \tag{2}$$

Each measure will of course be linked to a cost, measured in monetary values, denoted as $C(M_h^{t,l,a})$, $h = 1, \ldots, H_t^l$, $l = 1, \ldots, L$, $t = 1, \ldots, T$, $a = 1, \ldots, A$. Apart from the cost, the implementation of each measure will mean an effort, problems of various kind, the necessity to seek the access to certain people, to ask for permissions, uncertainty as far the effect of the measure is concerned, etc. - usually the bigger, the higher the level index is. Thus it is assumed that the effort depends on the level, not on the individual measures (again, it is only a simplifying assumption). It will be denoted as E_t, $t = 1, \ldots, T$ and evaluated by an expert in a fixed scale. Each single measure $M_h^{t,l,a}$, $h = 1, \ldots, H_t^l$, $l = 1, \ldots, L$, $a = 1, \ldots, A$ will be thus linked to effort E_t.

[2] In [6] there is a discussion showing that from the practical point of view all the formulae are in fact equivalent.

3 Model optimising the choice of countermeasures in project risk management

In the model we assume that the project manager wants, if possible, to achieve a small project overall risk level at the minimal cost. Also, the effort linked to the measures should be taken into account. Thus in fact three objective functions should be present in the model: the overall risk level, the cost of applying risk mitigation or elimination measures and the overall effort of applying those measures.

In order to solve this multicriteria problem, let us choose the easiest approach, where the level of all but one criteria has to be fixed, so that only one criteria formally remains as an objective function. In our opinion in most cases the decision maker will have to maintain a certain level of overall risk, and this level would be fixed beforehand. The decision maker would thus seek a way to achieve this level at the minimal cost. Also, he would like to minimize the effort, but probably rather as a secondary goal, trying to control it and be aware of it than to search for an absolute minimum. That is why the following model is proposed, but of course other formulations of this multicriteria problem would be possible too.

$$\sum_{\substack{h=1,\ldots,H_t^l,\, l=1,\ldots,L,\\ t=1,\ldots,T,\, a=1,\ldots,A}} C(M_h^{t,l,a}) \cdot x_h^{t,l,a} \to min \tag{3}$$

$$\sum_{\substack{l=1,\ldots,L,\, t=1,\ldots,T,\\ a=1,\ldots,A,\, k=1,\ldots,P_t^l}} \frac{\sum_{a=1}^{A} w_a(R_k^{t,l}) \cdot (1 - y_{k,a}^{t,l}) \cdot v_a(R_k^{t,l}) + \sum_{a=1}^{A} w_a(R_k^{t,l}) \cdot y_{k,a}^{t,l} \cdot m(v_a(R_k^{t,l}))}{\sum_{a=1}^{A} w_a(R_k^{t,l})} \leq RL \tag{4}$$

$$\sum_{k \in \{1,\ldots,P_t^l\} \text{ and } M(v_a(R_k^{t,l}))=h} y_{k,a}^{t,l} \leq M \cdot x_h^{t,l,a} \tag{5}$$

for each $h = 1, \ldots, H_t^l$, $l = 1, \ldots, L$, $t = 1, \ldots, T$, $a = 1, \ldots, A$

$$\sum_{\substack{h=1,\ldots,H_t^l,\, l=1,\ldots,L,\\ t=1,\ldots,T,\, a=1,\ldots,A}} E(M_h^{t,l,a}) \cdot x_h^{t,l,a} \leq EL \tag{6}$$

where $x_h^{t,l,a}$, $h = 1, \ldots, H_t^l$, $l = 1, \ldots, L$, $t = 1, \ldots, T$, $a = 1, \ldots, A$, will be equal to 1 if the corresponding measure should be applied and to 0 otherwise, $y_{k,a}^{t,l}$, $l = 1, \ldots, L$, $t = 1, \ldots, T$, $a = 1, \ldots, A$, $k = 1, \ldots, P_t^l$ will be binary variables fulfilling the constraint $y_{k,a}^{t,l} = 0$ (thus, being able to be eliminated from the model) if the value of the a-th attribute of risk $R_k^{t,l}$ would not be changed thanks to the corresponding countermeasure. The constant RL stands for the admissible overall project risk level, the constant EL for the admissible effort level linked to the selected countermeasures and the constant M for a sufficiently big number.

Actually, in practice most of the parameters in the above model may be impossible to determine in an exact way and may require a fuzzy formulation. Because of the limited scope of the paper, we assume, basing ourselves on [4, 5, 7, 9], only the risk attributes and the weights of the attributes to be fuzzy, as well the constants RL and EL and cost values will be crisp.

For the fuzzy modeling, let us use here the language from [9], where the authors consider three risk attributes (probability of occurrence, severity of consequences, early detection difficulty) whose values can take five fuzzy vales each (VH - very high, H - high, M - moderate, L

- low, VL - very low). The same 5 values are assigned in [9] to the attributes weights. The human language assumed in [9] at the background is as follows ($\widetilde{A} = (a_1, a_2, a_3)$ stands for the triangular fuzzy number with support $[a_1, a_3]$ and a_2 as the value with the membership degree 1):

Table 1: Linguistic scale for risk attributes, their weights and the effort linked to the countermeasures [9]

Linguistic term	**Corresponding triangular fuzzy number**
Very low (VL)	(0, 0, 0.25)
Low (L)	(0, 0.25, 0.5)
Moderate (M)	(0, 0.5, 0.75)
High	(0, 0.75, 1)
Very High (VH)	(0.75, 1, 1)

For RL and EL we will use fuzzy numbers $\widetilde{A} = (a_1, a_2, a_3)$ with $a_2 = a_3$. The decision maker might be supported in the choice of the fuzzy numbers for RL and EL by means of questions similar to the following ones:

- For RL: how many serious risks at the maximum would he accept, and how many serious risk he thinks have to be linked to each project at the minimum (as there are no projects without risk and accepting risks opens new possibilities)?
- For EL: similar questions, but regarding difficult countermeasures.

With fuzzy parameters denoted by $\widetilde{\ }$, we would have a mathematical programming problem with fuzzy parameters in constraint (4), which would become

$$\sum_{\substack{l=1,\ldots,L,\, t=1,\ldots,T,\\ a=1,\ldots,A,\, k=1,\ldots,P_t^l}} \frac{\sum_{a=1}^{A} \widetilde{w}_a(R_k^{t,l}) \cdot (1 - y_{k,a}^{t,l}) \cdot \widetilde{v}_a(R_k^{t,l}) + \sum_{a=1}^{A} \widetilde{w}_a(R_k^{t,l}) \cdot y_{k,a}^{t,l} \cdot \widetilde{m}(v_a(R_k^{t,l}))}{\sum_{a=1}^{A} \widetilde{w}_a(R_k^{t,l})} \leq \widetilde{RL} \quad (7)$$

The rest of model (3) - (6) will remain unchanged.

Constraint (7) is not unequivocal. It may be interpreted in many ways, according to the way the decision maker chooses to compare fuzzy numbers. Also, the fuzzy numbers on the left hand side have to be multiplied with a constant and added one to another.

Let us thus discuss shortly basic arithmetical operations on fuzzy numbers and the simplest approaches of comparing them. Let us thus consider two triangular fuzzy numbers $\widetilde{A}$ and $\widetilde{B}$, based, respectively, on the following triples of crisp numbers: (a_1, a_2, a_3) and (b_1, b_2, b_3). Then we have:

- $\widetilde{A} + \widetilde{B} = (a_1 + b_1, a_2 + b_2, a_3 + b_3)$ (8)
- $\widetilde{A} \cdot \widetilde{B} = (a_1 b_1, a_2 b_2, a_3 b_3)$ if all the parameters are positive (9)
- For each crisp number r we have: $r \cdot \widetilde{A} = \widetilde{A} \cdot r = (ra_1, ra_2, ra_3)$ (10)

If it comes to the inequality $\widetilde{A} \leq \widetilde{B}$, it may be interpreted depending on the pessimism/optimism degree of the decision maker. E.g. three interpretations listed below are possible, where the first one is the most pessimistic (it is most difficult for $\widetilde{A}$ to satisfy it) and the last one the most optimistic:

- $a_3 \leq b_1$ (11)
- $\frac{a_3+a_2}{2} \leq \frac{b_2+b_1}{2}$ (12)
- $a_2 \leq b_2$ (13)
- $\frac{a_2+a_1}{2} \leq \frac{b_3+b_2}{2}$ (14)
- $a_1 \leq b_3$ (15)

For other possibilities see e.g. [3]. Choosing (in cooperation with the decision maker) one of the approaches, we can turn constraint (7) into a crisp constraint.

4 Example

Let us consider a real world construction project, presented in detail in [8]. There are over 30 risks there and a similar number of countermeasures. Here let us consider only a selection of the data. Thus we have the following risks:

- on the project level, determinant time: $R_1^{1,1}$ - erroneous identification of soil;
- on the project level, determinant cost: $R_1^{1,2}$ - accidents during the construction;
- on the construction market level, determinant time: $R_1^{2,1}$ - non-availability of work force at the budgeted cost;
- on the construction market level, determinant cost: $R_1^{2,2}$ - non-availability of required equipment;
- on the national level, determinant cost: $R_1^{3,2}$ - unfavorable inflation.

The values of the risk attributes were assessed by the experts, as well as the weights of the attributes (we assume in the example that the weights are the same for all the risks).

Table 2: Risks attributes and their weights in the example

Risk	$\mathbf{R_1^{1,1}}$	$\mathbf{R_1^{1,2}}$	$\mathbf{R_1^{2,1}}$	$\mathbf{R_1^{2,2}}$	$\mathbf{R_1^{3,2}}$	**Weight of the attribute**
Probability of occurrence	M	VH	M	H	H	L
Consequences of occurrence	H	VH	H	M	VH	VH
Difficulty in early detection	VH	H	L	L	L	M
Difficulty in influencing the risk by countermeasures	L	L	H	H	VH	M

Five countermeasures have been identified, together with their cost and effort:

Table 3: Countermeasures, their cost, efforts and effects for the example

Countermeasure	**Risk and attribute influenced**	**Effect on the attribute**	**Cost of the countermeasure (in monetary units denoted $)**	**Effort linked to the countermeasure**
Having ready an alternative technology which can be used immediately when the soil problem is detected, $M_1^{1,1,2}$	$R_1^{1,1}$ consequences	H → L	40 $	L
Additional check of all the equipment and workers training, $M_1^{1,2,1}$	$R_1^{1,2}$, probability	VH→H	30 $	L
Training offered to pupils which are about to graduate a construction high school, $M_1^{2,1,2}$	$R_1^{2,1}$, probability	M→L	50 $	H
Renting or reserving the needed equipment already now, $M_1^{2,2,1}$	$R_1^{2,2}$, consequences	M→VL	40 $	H
Using a more stable currency in all the transactions, $M_1^{3,2,1}$	$R_1^{3,2}$, probability	H→L	20 $	VH

For the example we have:

- $l = 1, 2$ (we consider two project success determinants: time and cost)
- $t = 1, 2, 3$ (we have three levels: the project level, the construction market level, the national level);
- $a = 1, 2, 3, 4$ (we have four risk attributes corresponding to the rows of Table 2);
- $H_t^{l,a} = 1$ for the following triples (t, l, a), which can be read off from the rows of Table 3: $(1, 1, 2), (1, 2, 1), (2, 1, 1), (2, 2, 2), (3, 2, 1)$. Otherwise it is equal to 0;
- decision variables (binary) are as follows: $x_1^{1,1,2}, y_{1,2}^{1,1}, x_1^{1,2,1}, y_{1,1}^{1,2}, x_1^{2,1,2}, y_{1,1}^{2,1}, x_1^{2,2,2}, y_{1,2}^{2,2}$, $x_1^{3,2,2}, y_{1,1}^{3,2}$.

Let us assume $RL = EL = (2, 3, 3)$. We get then the following model (using values from Table 1,2 and 3 and formulae (8)-(10)):

$$40x_1^{1,1,2} + 30x_1^{1,2,1} + 50x_1^{2,1,2} + 40x_1^{2,2,2} + 20x_1^{3,2,2} \rightarrow min \tag{16}$$

$$\begin{aligned}A =&(0,75,1,1)(1-y_{1,2}^{1,1})(0,0,75,1)+(0,75,1,1)y_{1,2}^{1,1}(0,0,25,0,5)\\&+(0,0,25,0,5)(1-y_{1,1}^{1,2})(0,75,1,1)+(0,0,25,0,5)y_{1,1}^{1,2}(0,0,75,1)\\&+(0,0,25,0,5)(1-y_{1,1}^{2,1})(0,0,5,0,75)+(0,0,25,0,5)y_{1,1}^{2,1}(0,0,25,0,5)\\&+(0,75,1,1)(1-y_{1,2}^{2,2})(0,0,5,0,75)+(0,75,1,1)y_{1,2}^{2,2}(0,0,0,25)\\&+(0,0,25,0,5)(1-y_{1,1}^{3,2})(0,0,75,1)+(0,0,25,0,5)y_{1,1}^{3,2}(0,0,25,0,5)\\=&(0,0,75,1)(1-y_{1,2}^{1,1})+(0,0,25,0,5)y_{1,2}^{1,1}+(0,0,25,0,5)(1-y_{1,1}^{1,2})\\&+(0,0,1875,0,5)y_{1,1}^{1,2}+(0,0,125,0,375)(1-y_{1,1}^{2,1})+(0,0,0625,0,25)y_{1,1}^{2,1}\\&+(0,0,5,0,75)(1-y_{1,2}^{2,2})+(0,0,0,25)y_{1,2}^{2,2}+(0,0,1875,0,5)(1-y_{1,1}^{3,2})\\&+(0,0,0625,0,25)y_{1,1}^{3,2}\end{aligned} \tag{17}$$

$$B=(0,75,1,1)+(0,0,25,0,5)+(0,0,25,0,5)+(0,75,1,1)+(0,0,25,0,5)=(1,5,2,75,3,5) \tag{18}$$

$$\frac{A}{B}\leq(2,3,3) \tag{19}$$

$$y_{1,2}^{1,1}\leq Mx_1^{1,1,2}, y_{1,1}^{1,2}\leq Mx_1^{1,2,1}, y_{1,1}^{2,1}\leq Mx_1^{2,1,2}, y_{1,2}^{2,2}\leq Mx_1^{2,2,2}, y_{1,1}^{3,2}\leq Mx_1^{3,2,2} \tag{20}$$

$$\begin{aligned}&(0,0,25,0,5)x_1^{1,1,2}+(0,0,25,0,5)x_1^{1,2,1}+(0,0,75,1)x_1^{2,1,2}\\&+(0,0,75,1)x_1^{2,2,2}+(0,75,1,1)x_1^{3,2,2}\leq(2,3,3)\end{aligned} \tag{21}$$

Assuming the method (12) of comparing fuzzy numbers and combining (17), (18) i (19), we get the following crisp linear programming model:

$$40x_1^{1,1,2}+30x_1^{1,2,1}+50x_1^{2,1,2}+40x_1^{2,2,2}+20x_1^{3,2,2}\rightarrow min$$

$$\begin{aligned}&0,875(1-y_{1,2}^{1,1})+0,375y_{1,2}^{1,1}+0,375(1-y_{1,1}^{1,2})+0,34375y_{1,1}^{1,2}+0,25(1-y_{1,1}^{2,1})\\&+0,15625y_{1,1}^{2,1}+0,625(1-y_{1,2}^{2,2})+0,125y_{1,2}^{2,2}+0,34375(1-y_{1,1}^{3,2})+0,15625y_{1,1}^{3,2}\leq 2,5\end{aligned} \tag{22}$$

$$y_{1,2}^{1,1}\leq Mx_1^{1,1,2}, y_{1,1}^{1,2}\leq Mx_1^{1,2,1}, y_{1,1}^{2,1}\leq Mx_1^{2,1,2}, y_{1,2}^{2,2}\leq Mx_1^{2,2,2}, y_{1,1}^{3,2}\leq Mx_1^{3,2,2}$$

$$0,375x_1^{1,1,2}+0,375x_1^{1,2,1}+0,875x_1^{2,1,2}+0,875x_1^{2,2,2}+x_1^{3,2,2}\leq 2,5 \tag{23}$$

Solving model (16),(20),(22),(23) we get the information that measures $M_1^{1,1,2}$ and $M_1^{2,1,1}$ should be applied, at the cost of 80. However, if we change the weight of the risk attributes in the last column of Table 2, e.g. if the probability gets the weight VH, there is no solution and we have to accept a higher risk level: $RL=(3,4,4)$. In that case the cheapest way of attaining this risk level would be the application of measures $M_1^{1,2,1}$ and $M_1^{3,2,1}$.

5 Conclusions

A new method of selecting risk countermeasures to attain the desired risk level has been proposed. It allows to choose risk countermeasures taking into account such elements as various risk attributes and various weights assigned to them by the decision maker, various risk levels (the ones closer to the project, where we have more influence, and the ones further away), the monetary cost linked to the countermeasures as well as the immeasurable effort their application implies. The information required from the decision maker may be given in linguistic terms, modeled by fuzzy numbers. A real world example has been presented too.

Further research is necessary to verify the approach in other real world projects as well as to support the decision maker in the choice of the definition of fuzzy inequalities, which may be interpreted in many ways. The choice of the interpretation of fuzzy constraints has to correspond to the preferences of the decision maker, otherwise the model will not lead to a satisfying solution.

Bibliography

[1] Courtot H. (1988), *La gestion des risques dans les projets*, Ed. Economica, Paris.

[2] Camilieri E. (2011), *Project Success: Critical Factors and Behaviours*, Gower Publishing Company.

[3] Liu X. (2001), Measuring the satisfaction of constraints in fuzzy linear programming, *Fuzzy Sets and Systems*, 122(2): 263-275.

[4] Mandal S., Maiti J. (2013), Risk analysis using FMEA: Fuzzy similarity value and possibility theory based approach, *Expert Systems with Applications*, 41(7): 3527-3537.

[5] Meyer A., Zimmermann, H. (2011); Applications of fuzzy technology in business intelligence. *International Journal of Computers Communications & Control*, 6(3): 428-441.

[6] Norman T.L. (2010), *Risk Analysis and Security Countermeasure Selection*, CRC Press.

[7] Shang K., Hossen Z. (2013); Research Paper: Applying Fuzzy Logic to Risk Assessment and Decision-Making, Casualty Actuarial Society, Canadian Institute of Actuaries, Society of Actuaries, 1-59.

[8] Skorupka D. (2012), *Method of Construction Projects Risk Assessment*, Lambert Academic Publishing.

[9] Wang Y.-M., Chin K.-S., Poon G. K. K., Yang J.-B. (2009), Risk evaluation in failure mode using fuzzy weighted geometric mean, *Expert Systems with Applications*, 1195-1207.

[10] Zhang Y., Zhi-Ping P. (2013); An optimization method for selecting project risk response strategies, *International Journal of Project Management*, 32(3); 412-422.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 9(5):593-601, October, 2014.

Decision Model for Assessing Healthcare ICT Support Implications: User Perception

A.M. Oddershede, F.M. Cordova, R.A. Carrasco, F.J. Watkins

Astrid M. Oddershede, Felisa M. Córdova
University of Santiago of Chile
Industrial Engineering Department,Chile
astrid.oddershede@usach.cl
felisa.cordova@usach.cl

Rolando Carrasco
School of IEEE
Newcastle University/
Newcastle Upon Tyne, UK
r.carrasco@ncl.ac.uk

Francisco J. Watkins
University of Santiago of Chile,
Electrical Engineering Department, Chile
francisco.watkins@usach.cl

Abstract: This paper presents a multi-criteria decision model based upon user judgments to assist the evaluation process of an Information and Communication Technology (ICT) network system in health care to improve the quality of service (QoS). Measuring quality in health care services is not an easy task, as there are many competing goals involved, human, economic, communications technology, governmental and others. Integrating multiple criteria decision analysis (MCDA) methodology with modeling and simulation through Optimization Network Engineering Tool (OPNET) platform permit to characterize main ICT user and identify priority applications to examine network QoS requirements and implications. The proposed approach permitted to identify the main users, to elaborate a profile and characterization of the ICT support requirements according to their main daily task in answer to a service requirement. The results generate evidence related to the important factors effecting quality in hospital requirement as availability of services and the need for ubiquitous access to integrated information. The stakeholder interface perception and resources for ICT network support are investigated through a case study for Chilean hospitals.
Keywords: MCDA, Decision support, User perception, ICT Healthcare.

1 Introduction

Information and communications technology appears as an emerging concept in health care undertaking an important role for healthcare-related activities [3]. An information technology system provides stakeholders with several applications to support their duties in clinic care, medical research and administrative issues. These applications rely on the ICT network infrastructure and its performance. An ICT system should be a facilitator for health care users since they need to access all types of data existing on all types of systems. There is evidence [7], [9] that an ICT network system implementation generates an effect on the service and health care providers. In this sense it has become important to consider, [1], [2] evaluation mechanisms for ICT healthcare support and applications. The Health ICT design, implementation and its management has to consider risk for quality and efficiency on patient care and be in agreement with the type of user needs that could be attained through user profiling and requirements analysis.

The study is concerned with the development of an exploratory assessment model incorporating empirical data collected from the main users of the health network system.The model will be of assistance to find out user perception of quality of the service related to the communications system in a healthcare institution, to identify critical areas for QoS of each user type, to provide a decision making tool as a guidance to analyse and evaluate a networked system for health related activities, to compare the different requirements and to enable trade-offs in accordance to the institution necessity. With the purpose to recognize the properties of the system that could subsequently affect the degree of satisfaction with the ICT system in unanimity with user activities, a multi- criteria approach is utilised, modelling with AHP [13]. The obtained data allows generating user profiles and applications profiles to support the design of models for evaluating ICT healthcare network QoS through the Optimization Network Engineering Tool (OPNET) simulation platform to examine the network behaviour and performance.

Section 2 introduces the proposed evaluation approach. Section 3 gives notions of the applied AHP method. Following the case study description and its results is presented in section 4 providing information that is not currently available. In section 5, the conclusions are provided.

2 Healthcare ICT Network System Evaluation Approach

The evaluation approach considers integrating user perception modelling through AHP with network modelling and simulation to observe the network and examine performance. QoS offered by a particular network could be established by technical parameters that can be measured objectively. However, it is a difficult task to find a set of universal parameters for every type of service because there are many and dissimilar parameters involved in the performance evaluation. The QoS technical metrics related to each attribute has to be defined together with the applications profile. These profiles are influenced by healthcare user requests for developing a specific task. The user perception depends upon their needs, their precise applications and their expectations. Concerning stakeholder's perception, AHP modelling plays an important role generating network attributes of the system. It is of great utility to analyze which of the parameters would be relevant when considering the user perception for a determined service. The organization must then define a service level agreement (SLA) for their main applications. Once healthcare service application and user profiles are characterized it follows to model, design and simulate through Optimization Network Engineering Tool to examine the potential network performance.

This paper is focused on presenting the perception model related to the communications system and its results for recognizing priority health activities and distinguishing the critical IT network support resources that could lead to the improvement of the service. The results will reveal key QoS parameters that generate user and activities profiles to support the assessment of communications resources of Local Area Networks (LAN) in a hospital.

3 The Analytic Hierarchy Process Methodology

The AHP involves decision-makers (DM) in breaking down a decision into smaller parts, proceeding from the goal to criteria to sub-criteria down to the alternative courses of action. DMs then make pair-wise comparison judgements throughout the hierarchy to arrive at overall priorities for the alternatives [13]. This approach provides the structure and the mathematics to support decision-makers make rational decisions. The basic principles of AHP are: Hierarchy representation and decomposition, which is a representation of a complex problem in a multilevel structure whose first level is the goal followed successively by levels of factors, criteria, and sub criteria, and so on down to a bottom level of alternatives. Figure 1 shows an illustration of a

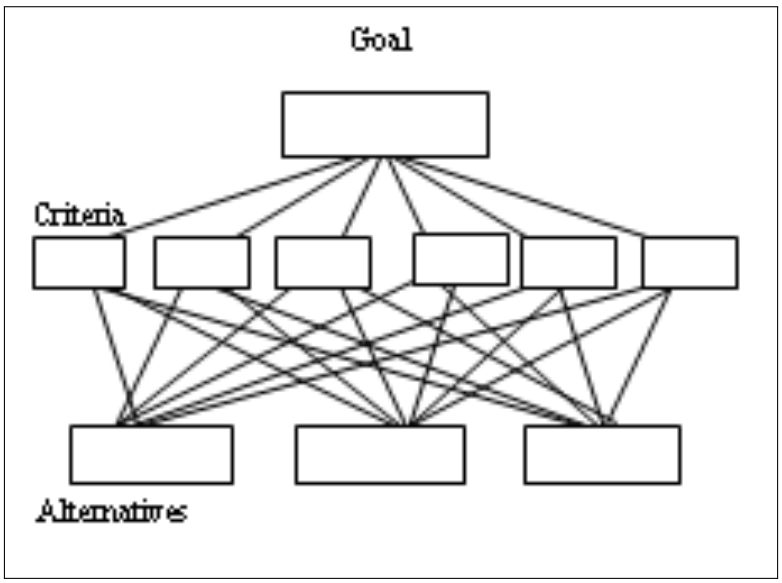


Figure 1: Generic Decomposition of a Problem into a Hierarchy

simple three level hierarchy. The object of a hierarchy is to assess the impact of the elements of a higher level on those of a lower level or alternatively the contribution of elements in the lower level to the importance or fulfillment of the elements in the level above. This type of assessment is usually made by paired comparisons responding to an appropriately posed question eliciting the judgement. The mathematical definition of a hierarchy is given in Saaty's Book [13].

Setting priorities in a hierarchy requires that we perform measurements throughout the structure. We must then synthesize these measurements to obtain priorities for the bottom level alternatives. The AHP is based on ranking activities in terms of relative ratio scales. In the paired comparison approach of the AHP, one estimates ratios by using a fundamental scale of absolute numbers in comparing two alternatives with respect to an attribute and one uses the smaller value as the unit for that attribute. To estimate the larger one as a multiple of that unit, assign to it an absolute number from a fundamental scale shown in table 1.

Table 1: Saaty's Fundamental Scale

Importance Intensity	**Definition**
1	Equal Importance
3	Moderate Importance
5	Strong Importance
7	Very strong or demonstrated Importance
9	Extreme importance
Reciprocals of above	If activity i has one of the above nonzero numbers assigned to it when compared with activity j, then j has the reciprocal value when compared with i

This process is done for every pair. Thus, instead of assigning two numbers w_i and w_j and forming the ratio w_i/w_j we assign a single number drawn from the fundamental $1-9$ scale to represent the ratio $(w_i/w_j) : 1$. The absolute number from the scale is an approximation to the ratio w_i/w_j. The derived scale tells us what the w_i and w_j are. Let W be a matrix (1)

whose row elements are ratios of the measurements *wi* of each of n items with respect to all others.

$$W = \begin{bmatrix} w_1/w_1 & \cdots & w_1/w_n \\ w_2/w_1 & \cdots & w_2/w_n \\ \vdots & \ddots & \vdots \\ w_n/w_1 & \cdots & w_n/w_n \end{bmatrix} \tag{1}$$

A number in the matrix is a dominance judgment. A judgment of 1.0 means that two activities contribute equally to the objective or goal, a judgment of 3.0 means that slightly favour one activity over another or three times as much (if you are dealing with measurable), a judgment of 5.0 means that judgement strongly favour one activity over another, a judgment of 7 means that activity is strongly favoured over another; its dominance is demonstrated in practice and 9.0 means that the evidence favouring one activity over another is of the highest possible order of affirmation. You should group your elements into homogeneous clusters so that it is not necessary to use a number larger than 9. In this way, we can interpret all ratios as absolute numbers or dominance units. The AHP provides guidelines for a test of consistency of judgments to ensure that elements are grouped logically and ranked consistently according to a logical criterion. In general, the ratio should be in the neighborhood of 0.10 [13]. Too great a departure from the perfectly consistent value indicates a need to improve the judgments or to restructure the hierarchy.

4 The Case Study and AHP application

The stakeholder interface perception and resources for ICT network support are investigated through a case study for Chilean hospitals. A pilot study has been carried out collecting data from health Institutions in Chile (private, public, regional) to examine ICT infrastructure, ICT network provision and stakeholders perception related to ICT network system. An AHP model is constructed to determine user perspective related to the ICT support importance in developing their work. For this study, we classified the main ICT network system users into three groups: those who develop activities in clinic care, who would make use of ICT to deliver a service (Physician, nurses, paramedics, etc.); the medical research group, who develop health research, collecting disease statistic and/or investigate new drugs and new devices and a third group integrated by users who perform administrative activities, billing, products distribution, and inventory control or other connected. For this study patients were not considered since from previous work, [10] ICT support showed to have a lower impact on patient [1]. Figure 2 show the AHP process results. ICT support reflects the greatest impact on supplying clinical care service.

The next step is concentrated in finding out the relevant ICT application for each type of healthcare users that would support to perform a better service. Initial data was collected from 480 participants; following an expert team of 36 is grouped comprising representatives from each category from all of the three types of hospitals considered. The ICT applications to be supported by a server are: Email, Web Browsing (Http 1.1), File transfer, Database Access, File Print, Video Conferencing, and Voice.

A new AHP model is developed and processed. The pairwise results from the three group representatives indicated the *relative importance* of ICT system application for performing their

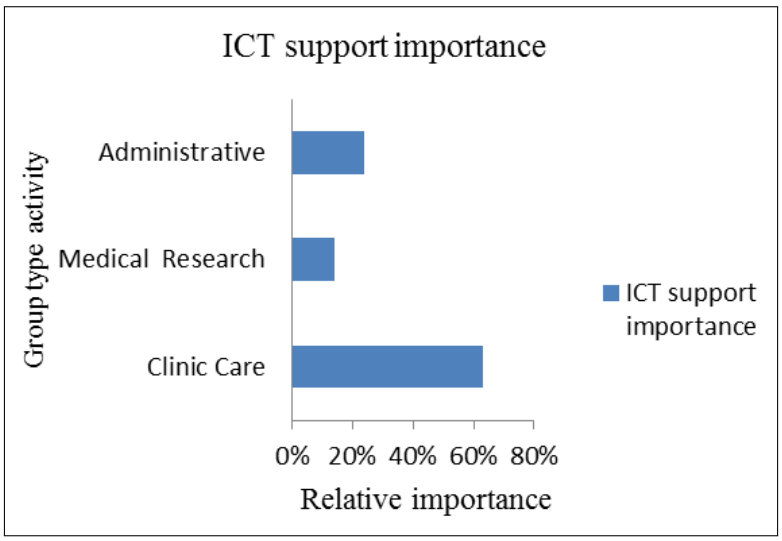


Figure 2: Relative ICT Support Importance for User Type Group

Table 2: Overall Relative Importance Priority for ICT Applications

Application	**Priority**
Data base access	51, 1
Web browsing	16, 5
FTP	2, 9
E mail	14, 5
Video and Voice	2, 9

activities as seen in table (2).

Though, when analysing separately, the clinical care group revealed a strong tendency, to rely on database applications to have access to patient records throughout email services. The research group indicated a strong interaction with Web Browsing application and data base application. This result would be in concordance to the nature of their work. While from administration group perspective, the preferences are for data base and file transfer protocol. The activities such as, delivering and obtaining test and exams results, within the institution implies interaction with database application. The applications relative importance according to ICT user group for each application is depicted in figure 3.

Regarding to current ICT applications usage in Healthcare Institutions there is a gap between current usages compared to what users declare important. From data collected and author's observations the application, e-mail, appears to be mostly employed and there is little usage of the others. Figure 4 shows the usage according to different health Institutions. In this sense there is work to be done, moreover, the results analysis suggests that health ICT network users expectations are that network will help to deliver a service with the required functionality on time and within the budget.

Consequently, a new AHP model is developed, to bring about information related to the most important network attributes to develop the applications. The essential QoS attributes to meet the ICT support requirement for each defined activity has to be established. It follows the

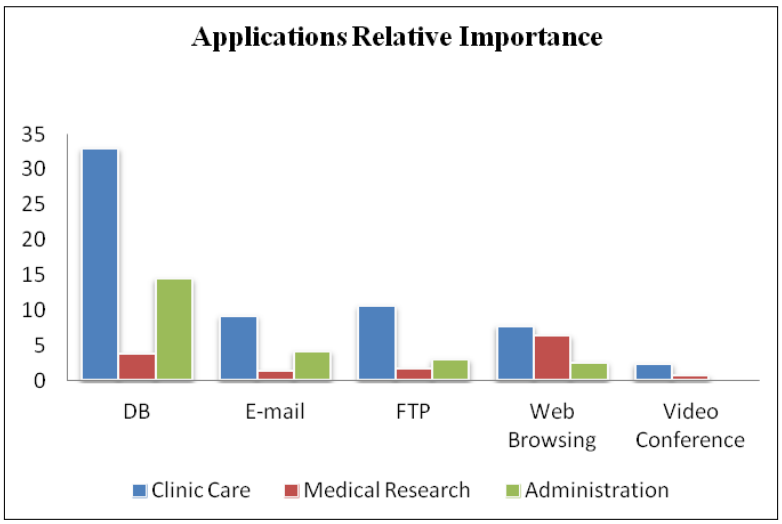


Figure 3: Relative ICT Support Importance for User Type Group

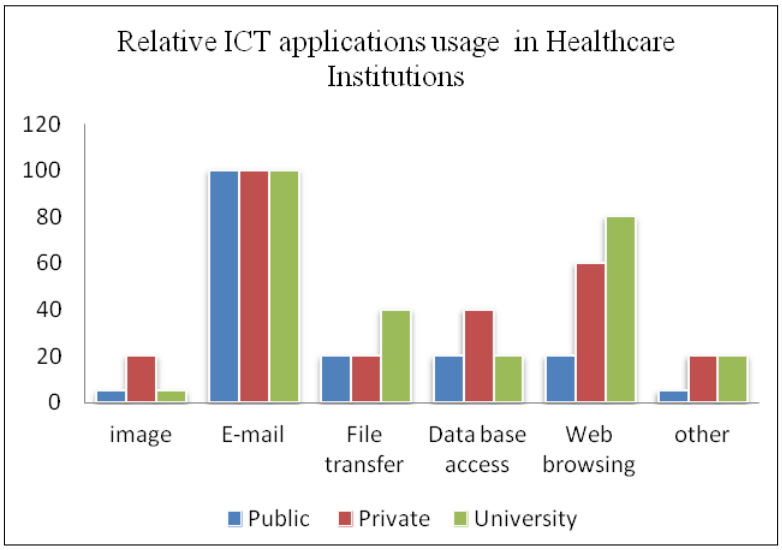


Figure 4: Applications Usage in Healthcare Institutions

process to determine the QoS attribute relative importance for applications according to user type expectation. This refers to the ICT support, user perceives about service satisfaction. For example: success in the connection, accessibility, velocity, etc. ICT system users expect that the network will help to deliver a service on time with the required functionality and within the budget.

Then the hierarchy structure should consider the attributes that would improve/ensure a better performance. The attributes considered are based on standard ISO [6] quality software model: Functionality, Efficiency, Reliability, Availability, and Serviceability. Even though, the five attributes are essential, it was possible to detect some differences in relation to the type of user. The two overall most important QoS attributes concomitant to the ICT applications in performing their health related activities system are *availability* and *reliability*. The attribute *availability* is most important for the group who develop task in Clinic care. In effect, clinic care professionals require having information on their patients including those elaborated by others, as, complementary tests results, at the moment and at the place of attendance. *Availability* of

ICT support is critical, when dealing with an emergency situation. Figure 5 shows the overall relative importance for attributes that the panel of experts indicated.

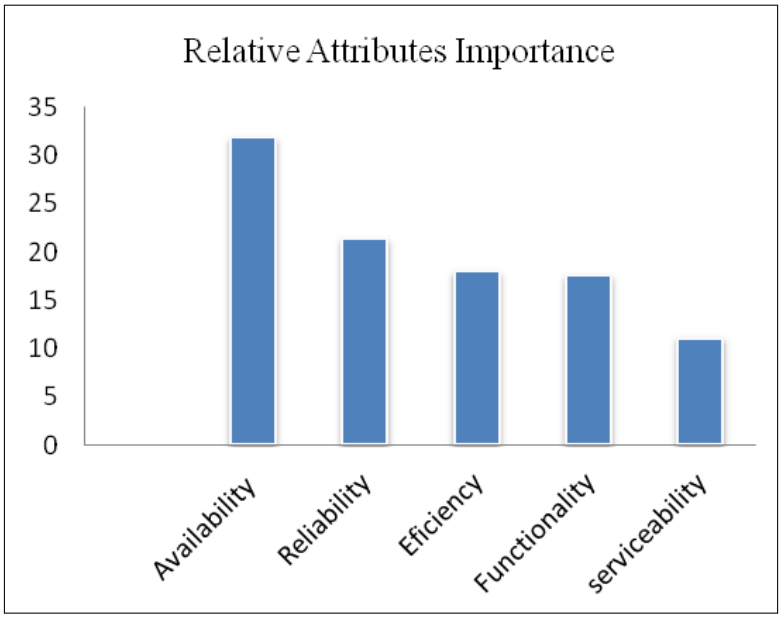


Figure 5: User relative importance for attributes

Through the AHP it was possible to identify applications priority and to characterize main ICT system users. The relative importance of ICT network support for each activity to deliver a better service is obtained. At this point, the QoS technical metrics that would guarantee a service related to each attribute are defined together with the applications profile. Technical requirement demanded for the different activities the group perform is shown in figure 6.

This information and data obtained by author's research and extensive collaboration with the IT network and management teams from a range of hospitals which differ in size and category as public, university, private is used for configuring profiles which will be the input for modelling and simulation to examine ICT network technical aspects and behaviour through OPNET methodology [11].
Then the next step is to set up profile applications according to each user type which is based on the AHP results obtained. According to OPNET methodology, initially a topology has to be selected from a hospital zone as a first approach to analyze technology infrastructure and network performance. Continuing with traffic configuration where every workstation will have a profile application consistent with the users' main role. Once applications and profiles are defined, then different scenarios are characterized for each study case to visualize how sensible network performance is, related to changes. Then, simulations are ran increasing the number of ICT network user, formulating scenarios, varying the number of users and/or varying the links (Ethernet connections) between workstations and switches, to obtain the point to point throughput in bits/secs. and utilization (link usage %). The Utilization percentage of link usage is expressed as:

$$\frac{Point - to - point.Throughput(bits/sec) * 100}{Ethernetconnection(Mbps)} \tag{2}$$

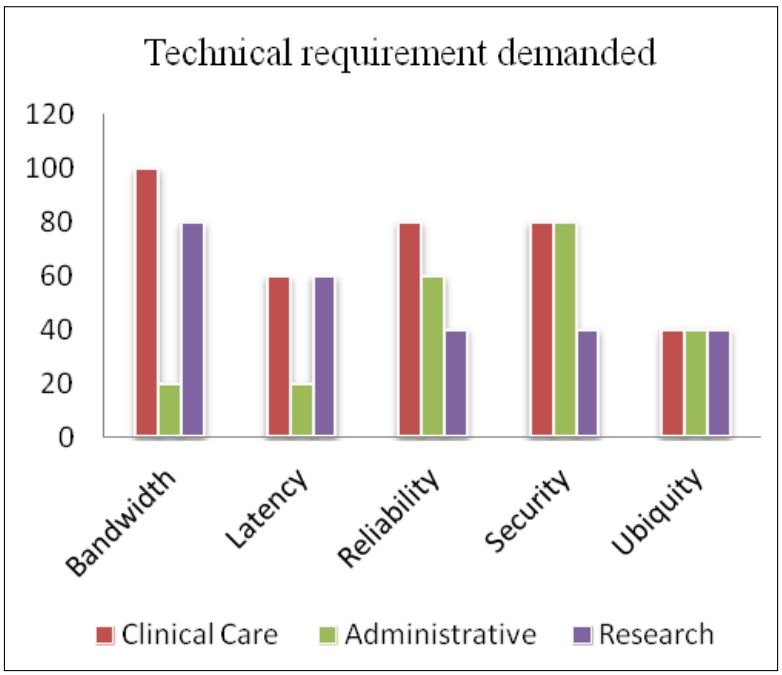


Figure 6: User Relative Technical Requirement Demanded

5 Conclusions

The existing competing objectives for improving quality of service in Health Institution increase complexity when analysing ICT support. The utilisation of a scientific multicriterial decision method AHP was beneficial for identifying high-priority requirements of an ICT system in health related activities.

The attribute for measuring quality in Clinic healthcare system, availability and the need for ubiquitous access to integrated information are considered most important. The combination of fixed and wireless network support can facilitate to obtain the timely information needed. This issue is critical in healthcare institution operation mainly for the clinic care group and work has to be done in this sense.

The MCDA approach allowed elaborating a profile and characterization of the ICT support requirements in healthcare service. The use of AHP modelling and empirical evaluation permits to capture human perspective. This allowed designing profiles characterization that would help to configure network traffic and different scenarios for simulating and evaluating network behaviour.

Acknowledgements

The authors are grateful to the Department of Industrial Engineering and DICYT of the University of Santiago of Chile for its support.

Bibliography

[1] Ammenwerth E., Nykanen P., Rigby M., Keizer N. (2013); *Clinical Decision Support Systems: Need for Evidence, Need for Evaluation*, Institute of Health Informatics, UMIT - 06.

[2] Ammenwerth E., Graber S., Herrmann G., Burkle T., Konig J. (2003); Evaluation of health information systems problems and challenges, *International Journal of Medical Informatics*, 71: 125-135.

[3] Bourret C. (2004); Data Concerns and Challenges in Health: Networks, Information Systems and Electronic Records, *Data Science Journal*, 3:96-113.

[4] Gao F., Ye X. (2002); A Hierarchical Trade-off Assessment Model and the Systematic Evaluation of Networked Systems, *Fast Abstract ISSRE*, Copyright Chillarege Press.

[5] Heath A., Carrasco R.(2001), Access techniques for 3G multimedia wireless packet switched networks: simulation using OPNETTM, *IEE/IEEE/BCS 6th International Symposium on Communication Theory and Applications (ISCTA01)*, Lancaster University, 15-20 Jul 2001, 1-29.

[6] ISO/IEC: ISO/IEC 9126-1(1997); Information Technology - Software Quality Characteristics and Metrics Part 1: Quality characteristics and sub-characteristics.

[7] Jaspers M.W., Smeulers M. Vermeulen H., Peute L.W. (2011), Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings, *Journal of the American Medical Informatics Association*, 18: 327-334.

[8] National Research Council. (2000); *Networking Health: Prescriptions for the Internet*, National Academic Press ISBN-10: 0-309-06843-6.

[9] Oddershede A. (2009); *Methodology to Evaluate QoS of ICT Networks for the Chilean National Health Service*, A thesis submitted to Newcastle University for the degree of Doctor of Philosophy, 2009.

[10] Oddershede A.M., Carrasco R.A (2007); Perception of Mobile Technology Provision in Health Service, Chapter book Global Mobile Commerce: Strategies, Implementation and Case Studies, edited by Dr. Wayne Huang, Dr. Y.L. Wang and Dr. John Day, Ohio University, USA.

[11] Opnet: User Manual (2004); `http://www.opnet.com/university\_program/teaching\_with\_opnet/textbooks\_and\_materials/materials/OPNET\_Modeler\_Manual.pdf`

[12] Xinjie Chang (1999); Network Simulations With Opnet, *Proc. of the 1999 Winter Simulation Conference*, P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds., Network Technology Research Centre School of EEE Nanyang Technological University, Singapore.

[13] Saaty, Thomas L. (2001); *Decision Making for Leaders*, Vol. II, AHP Series, RWS Publ., (new ed.), ISBN 0 - 9620317.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 9(5):602-609, October, 2014.

Optimal Routing Strategy Based on Specifying Shortest Path

F. Shao, B. Cheng

Fei Shao*, Binghua Cheng
School of Computer Engineering, Jinling Institute of Technology
No. 99, Hongjing Rd., Nanjing, 211169, JiangSu, P.R.China
*Corresponding author: shaofei@jit.edu.cn
cbh@jit.edu.cn

Abstract: How to enhance the transfer capacity of weighted networks is of great importance. The network transfer capacity, which is often evaluated by the critical packet generation rate, is proved to be inversely proportional to the highest node betweenness. By specifying the shortest path according to the different node characteristics, two different routing strategies are proposed to reduce the high node betweenness for the different node delivery capability schemes. Simulations on both computer-generated networks and real world networks show that our routing strategies can improve the network transfer capacity greatly. Especially, the greater the new added edge number is, the more efficient our routing strategies are.
Keywords: complex networks, weighted networks, routing strategy, betweenness.

1 Introduction

Due to the constantly growing significance of large scale communication networks such as the World Wide Web, the network transfer capacity has attracted an increasing attention. In those previous studies to improve transfer capacity and control traffic congestion on networks, some focus on making appropriate adjustments to the network topology structure [1] - [3] while others on finding optimal routing strategies [4] - [9]. Since the former is too expensive and too difficult to implement in some large-scale networks, most of previous works focused on effective routing strategies. Some of them are based on the global information: the shortest path routing strategy [4] which pass through the minimum number of nodes, the efficient path routing strategy [5] whose sum of node degrees is the minimum. Some others focus on local topological information since global information is usually unavailable in large-scale networks: the neighbor information [6], the next-nearest-neighbor information [7].

However, those aforementioned studies are mostly focused on the simplest unweighted networks with edges between nodes are represented by binary states according to whether the edges are present or not. In fact, the scientific collaboration networks [10], the cellular metabolism [11], the world-wide airport networks [12] and the Internet [13] have been proved to be weighted networks which are specified not only by its topology but also by the weight of the edges. Lots of models have been presented to describe weighted network among which the BBV weighted network model, coupled dynamical evolution of topology and weights, is most widely used.

In those most widely used traffic models in weighted network, the packets are transferred through the traditional shortest path [4], which is a path with the minimal number of nodes between arbitrary pairs of nodes, or the weighted shortest path [10, 14], where the distance between nodes is just the inverse of the weight of edge linked them. However, it is proved that these two routing strategies are not optimal for weighted networks [15]. Since there may be more than one traditional shortest path between some pairs of nodes, we can specify the best one to enhance the network transfer capacity.

This paper is organized as follows. In section 2 we describe the model and our routing strategy, followed by the experimental evaluations on the computer generated networks and real world network in section 3. The conclusions are given in section 4.

2 Models

2.1 Network Model

The BBV network can be completely described by a weighted adjacency matrix W whose elements w_{ij} denote the weight of the edge linking node i and j. The generation of the BBV network is based on two coupled mechanisms:

i Topological growth. Starting from the initial network with N_0 nodes which are fully connected by edges with assigned weight w_0, one new node is added at every time step. The new added node will be connected to m different nodes with equal weight w_0 for every edge and will choose nodes with large strength according to the probability $\prod_{n\to i} = s_i / \sum_l s_l$, where $s_i = \sum_j w_{ij}$ is the node strength.

ii Weight dynamics. The weight of each new added edge is initially set to a given value w_0 which is often set to 1 for simplicity. But the adding of edge connecting to node i will result in increasing the weight of the other edges linked to node i which is proportional to the edge weights. If the total increase is δ (we will focus on the simplest form: $\delta_i = \delta$), we can get

$$w_{ij} = w_{ij} + \Delta w_{ij} = w_{ij} + \delta * \frac{w_{ij}}{s_i} \tag{1}$$

This will yield the strength increase of node i as:

$$s_i = s_i + \delta + w_0 \tag{2}$$

The degree distribution of BBV network $P(k) \propto k^{-\gamma_k}$ and the strength distribution $P(s) \propto s^{-\gamma_s}$ yield scale-free properties with the same exponent [12], [16]- [18]:

$$\gamma_k = \gamma_s = \frac{4\gamma + 3}{2\gamma + 1} = 2 + \frac{1}{2\gamma + 1} \tag{3}$$

2.2 Traffic Model

The traffic model can be described as follows:

i All nodes can create packets with addresses of destination, receive packets from other nodes, and forward the packets to their destinations.

ii At each time step, there are R packets generated in the network, with randomly chosen sources and destinations. Once a packet is created, it is placed at the end of the queue if the node already has several packets waiting to be forwarded to their destinations.

iii At each time step, the first C_i packets at the top of the queue of node i, if it has more than C_i packets in its queue, are forwarded one step toward their destinations and placed at the end of the queues of the selected nodes. Otherwise, all packets in the queue are forwarded one step. This procedure applies to all nodes at every time step.

iv A packet, upon reaching its destination, is removed from the system.

In our model, three node delivery capability schemes are considered: (i) each node has the same packet delivery capability ($C_i = 1$, CON stands for this scheme); (ii) the node delivery capacity is considered to be proportional to the node strength s_i ($C_i = s_i/ < s >$, STR stands for this scheme); (iii) the node delivery capacity is considered to be proportional to the node degree k_i ($C_i = k_i/ < k >$, DEG stands for this scheme). To compare the overall transfer capacity, we normalize the delivery capability to keep the total node delivery capability of the whole network is equal to the node number n in three situations the same. When R is increased from 0 to ∞, two phases will be observed: free flow and congested phase. For $R < R_c$, the numbers of created and forwarded packets are balanced, resulting in a steady free flow of traffic. For $R > R_c$, traffic congestion occurs due to the fact that packet delivery capacity of node is limited. The phase transition from the former to the latter occurred at the critical packet generation rate R_c. We focus on the critical value R_c which can best reflect the transfer capacity of a network.

We utilize the betweenness b_i [19] to estimate the traffic passing through a node i under a given routing strategy:

$$b_i = \sum_{s,t} \frac{\sigma(s,i,t)}{\sigma(s,t)} \tag{4}$$

where $\sigma(s,i,t)$ is the number of paths under the given routing strategy between nodes s and t that pass through node i and $\sigma(s,t)$ is the total number of paths under the given routing strategy between s and t and the sum is over all pairs s,t of all distinct nodes.

The probability a packet will pass through the node i is $b_i/\sum_{j=1}^{n} b_j$, and therefore the average number of packets that the node i will receive at each time step is. When the number of incoming packets is equal to or larger than the outgoing packets at the node i, $Rb_i/(n(n-1)) \geq C_i$, traffic congestion will occur. So the critical packet generation rate R_c is

$$R_c = min\left(\frac{C_i * n * (n-1)}{b_i}\right) = n * (n-1) * min(C_i/b_i) \tag{5}$$

2.3 Routing Strategy

Enlightened by the efficient path [5], we also define$P_{i\to j}$ as the path between node i and j which pass through the nodes sequence $x_0(=i), x_1, x_2, \cdots, x_{n-1}, x_n(=j)$. However we define

$$F(P_{i\to j}, \alpha) = \sum_{i=0}^{n-1} w_{ij}^{\alpha} \tag{6}$$

In our routing strategies, we specify the path between i and j as the one makes $F(P_{i\to j}, \alpha)$ minimum under a given tunable parameter α. When α is -1, the specified routing strategy is the same as the weighted shortest path routing strategy [10, 14] (WSH stands for this routing strategy). When α is 0, the specified routing strategy is the same as the traditional dijkstra shortest path routing strategy [4] which pass through the minimum amount of nodes (SHT stands for this routing strategy). As we mentioned above, there may be more than one shortest path between some nodes. We calculate the sum of strengths s_i of all nodes on each shortest path, and select the one with minimum sum as our specified path. SSS stands for this routing strategy and SSD for the minimum sum of degrees k_i. The definition of our routing strategies (the SSS routing strategy and the SSD routing strategy) is shown in Tab.1.

Table 1: Definition of our routing strategies (SSS and SSD)

	WSH	SHT	SSS	SSD
$F\left(P_{i\to j},\alpha\right)$	$min\left(\sum_{i=0}^{n-1} w_{ij}^{-1}\right)$	$min\left(\sum_{i=0}^{n-1} w_{ij}^{0}\right)$	SHT&$min\left(\sum_{i=0}^{n-1} s_i\right)$	SHT&$min\left(\sum_{i=0}^{n-1} k_i\right)$

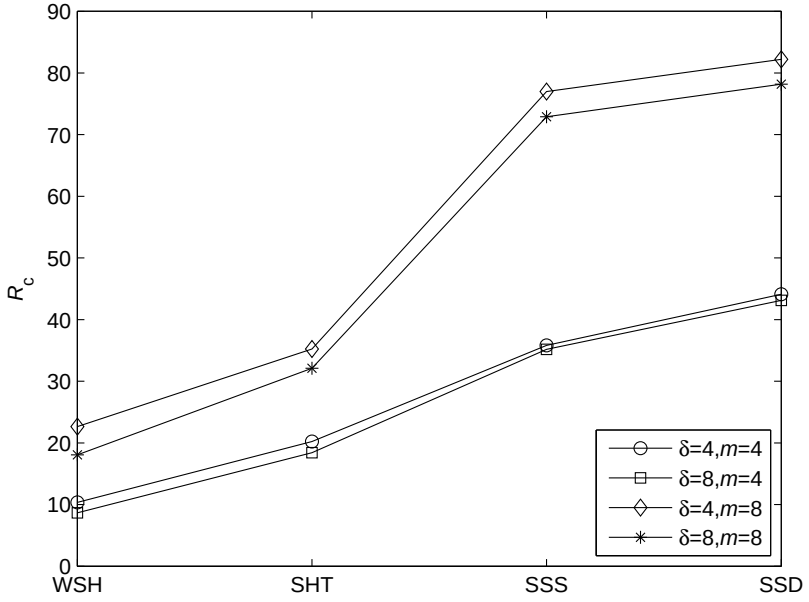


Figure 1: R_c of different routing strategies. BBV network with $n = 200$ and $w_0 = 1, C_i = 1$

3 Results and Discussion

To obtain the critical packet generation rate R_c in simulations, we use the order parameter[1]:

$$\eta = \lim_{t\to\infty} \frac{\langle \Delta\Theta \rangle}{R\Delta t} \tag{7}$$

where $\Delta\Theta = \Theta(t + \Delta t) - \Theta$, with $\langle \cdots \rangle$ indicating average over time windows of width Δt , and $\Theta(t)$ is the total number of packets in the network at time t. At the early stage, when R is very small, the generated packets can be delivered, $\langle \Delta\Theta \rangle$ is less than zero and so is η. Where η is greater than zero, we can obtain the critical packet generation rate R_c.

In figure 1, we plot the critical packet generation rate R_c of different routing strategies in a BBV network with $n = 200$ and $\omega_0 = 1$. (For every network, 10 instances are generated and for each instance, we run 10 simulations. The results are the average over all the simulations.)

Fig.1 shows that when each node has the same packet delivery capability, the WSH routing strategy is the most sensible to traffic congestion. The SHT routing strategy is better than WSH, and the SSD routing strategy has the maximum transfer capacity. Moreover, the greater the new added edge number m is, the more efficient our routing strategies are. Take networks with $\delta = 4, m = 4$ and $\delta = 4, m = 8$ for example, the SHT, SSS and SSD routing strategies can enhance the critical packet generation rate R_c 95.72%, 246.53 %, 326.75% than the WSH routing strategy correspondingly in network with $\delta = 4, m = 4$ while 112.95%, 306.81%, 398.81%

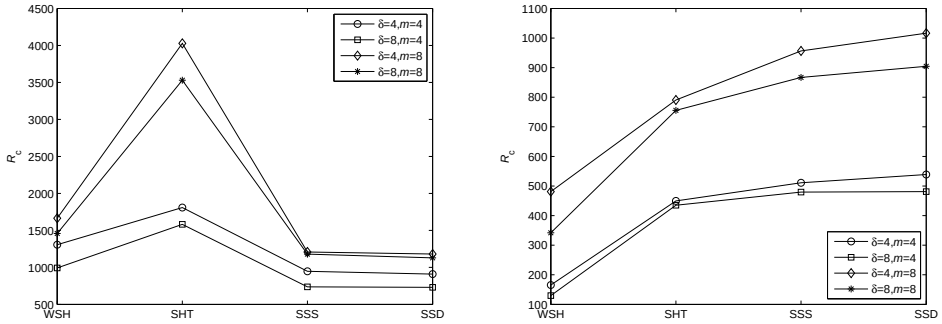


Figure 2: R_c of different routing strategies. BBV network with $n = 200$ and $w_0 = 1$ (a) $C_i = s_i/ < s >$ (b) $C_i = k_i/ < k >$

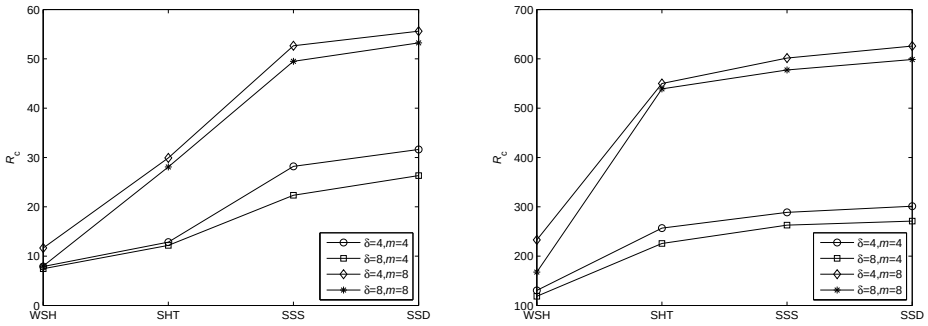


Figure 3: R_c of different routing strategies. BBV network with $n = 100$ and $w_0 = 1$ (a) $C_i = 1$ (b) $C_i = k_i/ < k >$

in network with $\delta = 4, m = 8$. When the new added edge number m is increased, there are more edges in the network and there might be more traditional dijkstra shortest paths between nodes consequently. That is why our routing strategies are more efficient with larger parameter m.

Then we turn to the other two schemes: the node delivery capacity is proportional to the node strength s_i and the node degree k_i. Simulation results are shown in Fig.2(a) and Fig.2(b) correspondingly.

Fig.2(a) presents that when the node delivery capacity is considered to be proportional to the node strength, the SHT routing strategy is the most effective while the SSD routing strategy has the largest R_c when the node delivery capacity is considered to be proportional to the node degree as shown in Fig.2(b). Fig.2(a) shows that our routing strategies do not work well in STR scheme. In DEG scheme, the SHT routing strategy is also better than WSH, and the SSD routing strategy still has the maximum transfer capacity. However, the gap among the SHT and SSS and SSD routing strategies is narrowed. And our routing strategies are more efficient with the greater new added edge number.

Then we check the impact of the node number n on our routing strategies. We test our routing strategies on BBV weighted networks with $n = 100$ nodes to achieve the simulation results of CON scheme and DEG scheme as shown in Fig. 3.

Fig.3(a) and Fig.3(b) display the influence of node number our routing strategies. We can

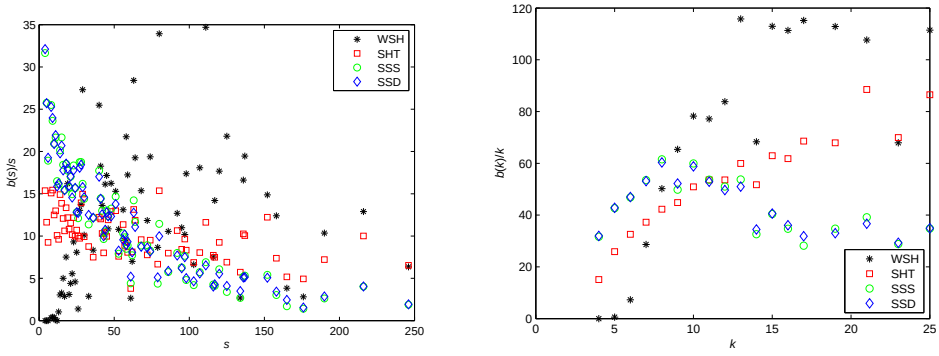


Figure 4: Betweenness per node. BBV network with $n = 200$, $\delta = 4, m = 4$ and $w_0 = 1$ (a) $C_i = s_i / < s >$ (b) $C_i = k_i / < k >$

discover that the SSD route is still the best way to enhance the critical packet generation rate. And by comparing Fig.3(a) with Fig.1 and Fig.3(b) with Fig.2(b), we can discover that the node number n has a little effect on the transfer capacity.

To achieve heuristic explanation for the routing strategies corresponding to the highest transfer capacity, we investigate the betweenness distribution on the network as presented in Fig.4.

The betweenness normalized by the strength of STR scheme is shown in Fig.4(a) and normalized by the degree of DEG scheme shown in Fig.4(b). In both figures, the load of the most effective routing strategy is distributed more evenly than the other three. In Fig.4(a), the betweenness divide by the strength of the SHT routing strategy is relatively flat which means the node with higher strength forward more packets. And in Fig.4(b), the node with higher degree forward more packets while using the SSD routing strategy. It is obvious that the traffic load under the SHT routing strategy for STR scheme and under the SSD routing strategy for DEG scheme are distributed evenly to the nodes according to their strength and degree correspondingly.

Those routing strategies which elongate the average path length L_{AVE} unnecessarily may not be efficient for network communications. Thus it is of great importance for a routing strategy to maintain the small-world phenomenon, i.e. $L_{AVE} \propto \ln n$. In our routing strategies, all the paths are the shortest path between arbitrary nodes which means the small-world phenomenon is still maintained in our routing strategies.

Finally, we test our routing strategies for three schemes on real world network. We choose the USAir 97 network (network of direct flight connections between US airports for the year 1997, http://vlado.fmf.uni-lj.si/pub/networks/data/) with 332 nodes and 2126 edges. Simulation results are shown in Tab.2.

Table 2: R_c of different routing strategies of USA airport network

	WSH	SHT	SSS	SSD
CON	3.49	4.86	5.85	5.87
STR	2.24	2.26	2.24	2.24
DEG	83.69	161.96	169.03	169.90

From Table 2 we can discover that in CON scheme, the SSD routing strategy has the maximum transfer capacity which is only a bit higher than the SSS routing strategy. The WSH

routing strategy has the lowest transfer capacity. It means when each node has the same packet delivery capability, our SSS routing strategy and our SSD routing strategy can enhance the transfer capacity in real world network. And in STR scheme, the SHT routing strategy has the maximum transfer capacity which is also the same as the computer generated weighted network. When it turns to the DEG scheme, the SSS routing strategy and the SSD routing strategy also achieve better results than the traditional routing strategy. In a word, the SSS routing strategy and the SSD routing strategy also works well in the real world network in the CON scheme and in the DEG scheme.

4 Conclusions

Considering the different node delivery capability, this paper has proposed two novel routing strategies to enhance the network transfer capacity in weighted networks. The characteristic of our strategy is to specify the shortest path according to three kinds of different node delivery capability schemes. The simulation shows that when each node has the same packet delivery capability, we can select the path with the minimal number of nodes and with minimum sum of node degree. And this routing path is also optimal in the scheme which the node delivery capacity is considered to be proportional to the node degree. When the node delivery capacity is considered to be proportional to the node strength, our routing strategies do not work. It is worth mentioning that our routing strategies are more efficient with the more new added edge. At last, we apply our routing strategies on the USAir 97 network to show the validity of our routing strategies on real world network. Moreover, the above-mentioned research may throw lights on designing better communication protocols.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (Grants No. 61373136 and 61375121), the Natural Science Foundation of Jiangsu Province, China (Grant No. BK2012082), the Research Foundation of Jinling Institute of Technology (Grant No. JIT-B-201406) and sponsored by Qing Lan Project. The author also gratefully acknowledges the helpful comments and suggestions of the reviewers, which have improved the presentation.

Bibliography

[1] Arenas A. et al (2001); Communication in networks with hierarchical branching, *Physical Review Letters*, ISSN 0031-9007. 86(14): 3196-3199.

[2] Guimera R. et al (2002); Optimal network topologies for local search with congestion, *Physical Review Letters*, ISSN 0031-9007, DOI://dx.doi.org/10.1103/PhysRevLett.89.248701.

[3] Zhang, G.Q. (2010); On cost-effective communication network designing. *Europhysics Letters*, ISSN 0295-5075. 89(3): 38003.

[4] Dijkstra E.W. (1959); A note on two problems in connexion with graphs. *Numerische Mathematik*, ISSN 0029-599X, 1(1): 269-271.

[5] Yan G. et al (2006); Efficient routing on complex networks. *Physical Review E*, ISSN 1539-3755. 73(4): 046108.

[6] Wang W.X. et al. (2006); Traffic dynamics based on local routing protocol on a scale-free network. *Physical Review E*, ISSN 1539-3755, 73(2): 026111.

[7] Yin C.Y. et al (2006); Traffic dynamics based on an efficient routing strategy on scale free networks, *The European Physical Journal B*, ISSN 1434-6028. 49(2): 205-211.

[8] Toroczkai Z.; Bassler K.E. (2004); Network dynamics: Jamming is limited in scale-free systems. *Nature*, 428, 716 (15 April 2004), doi:10.1038/428716a.

[9] Wu Y.H. et al (2013); Performance Analysis of Epidemic Routing in Delay Tolerant Networks with Overlapping Communities and Selfish Nodes, *International Journal of Computers Communications & Control*, ISSN 1841-9844, 8(5): 744-753.

[10] Newman M.E.J. (2001); Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, ISSN 1063-651X. 64(1): 016132.

[11] Almaas E. et al (2004); Global organization of metabolic fluxes in the bacterium Escherichia coli, *Nature*, ISSN 0028-0836. 427(6977): 839-843.

[12] Barrat A. et al (2004); The architecture of complex weighted networks. *PNAS*, ISSN 0027-8424. 101(11): 3747-3752.

[13] Pastor-Satorras R.; Vespignani A. (2007); *Evolution and structure of the Internet: A statistical physics approach*, Cambridge University Press, ISBN 9780521714778.

[14] Brandes U. (2001); A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, ISSN 0022-250X, 25(2): 163-177.

[15] Shao, F. (2013); Optimal Transport on Weighted Networks for Different Node Delivery Capability Schemes, *The Scientific World Journal*, http://dx.doi.org/10.1155/2013/378083.

[16] Barrat A. et al (2004); Modeling the evolution of weighted networks. *Physical Review E*, ISSN 1539-3755. 70(6): 066149.

[17] Barrat, A. et al (2004); Weighted evolving networks: coupling topology and weight dynamics. *Physical Review Letters*, ISSN 1079-7114. 92(22): 228701.

[18] Barthelemy M. et al (2005); Characterization and modeling of weighted networks. *Physica A*: Statistical Mechanics and its Applications, ISSN 0378-4371. 346(1-2): 34-43.

[19] Freeman L.C. (1977); A set of measures of centrality based on betweenness. *Sociometry*, ISSN 0038-0431, 40(1): 35-41.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 9(5):610-622, October, 2014.

Flexible Service Oriented Network Architecture for Wireless Sensor Networks

A.P. Singh, O.P. Vyas, S. Varma

Akhilendra Pratap Singh, O.P. Vyas, Shirshu Varma
Department of Information Technology
Indian Institute of Infomation Technology
Allahabad,211012, U.P., India
*Corresponding author: akhil121282@gmail.com
opvyas@iiita.ac.in, shirshu@iiita.ac.in

Abstract: Wireless Sensor Network (WSN) is a combination of homogeneous and heterogeneous sensor nodes which are physically deployed at different places. Heterogeneous WSM can be characterized with different parameters such as hardware resources, application, services, software platform and network. WSN are powerful technology, which is used to build up many applications in the era of technological development. These developments demand homogeneous or heterogeneous services as per the user requirements at applications level. The requirement results in increasing the number of services over the network and hence the number of complexities. Complexities of wireless sensor networks are increasing due to the lack of adoption of new services, new protocols and interoperability between heterogeneous services with common communication architecture. This paper deals with these issues of wireless sensor networks, elaborating the need of generalized communication architecture for applications, developers and users, proposing generalized Flexible Service Oriented Network Architecture (FSONA) to solve above issues along with the detailed functionality of the proposed architecture.
Keywords: Wireless Sensor Networks (WSN), Flexible Service Oriented Network Architecture (FSONA). middleware architecture, interoperability, localization.

1 Introduction

In the modern epoch wireless sensor networks are increasing rapidly in current environment for different application such as health monitoring, target tracking, mobility of object, petroleum industries, pressure measurement at various levels, environmental change etc. Sensors are small devices which are used to compute the environmental changes according to application. These devices have been developed from organizations and vendors, due to this reason hardware platform, software platform and specification can be dissimilar. Heterogeneous sensor network introduces a major issue to sensor node with its various different specification, failure prone Sensors and operation in the current dynamic environment. Failure prone sensors may affect the performance of the network. Therefore Multi objective method enhances the performance of task and network life time after detecting the failure prone sensor nodes [6]. So there is requirement to identify the failure prone nodes.

To provide a common interface between the sensor node and its requirement for sensor network, middleware architecture has been evolved which is capable to tackle the heterogeneous environment without changing the specification and its platform. Sensor middleware can be defined as a standard interface which is useful for communication and requirements of sensor, users in terms of services without affecting the platform [16]. Middleware architecture is able to manage the requirements of node in the form of request, storage of data and constraints like when, how the data and service are needed for the node operation. Middleware architecture supports a range of machine which have different configuration and provides compatibility for communication among the network and client. Wireless sensor network has the various challenging issues

which can be handled with service oriented computing model [5]. Service oriented computing model is based on the approach of service oriented architecture. Service oriented computing model intends to provide better service availability, communication protocols and accessibility of services in efficient way without knowing low level detail of hardware implementation and infrastructures. Service oriented architecture has various advantages over traditional approach such as reusability, flexibility, interoperability and loose coupling which is beneficial to develop middleware architecture.

Service oriented computing method covers the functional and non functional requirements of the application in the current dynamic environment. Service oriented middleware [20] provides a connection between various components and assist for communication over a multiple channel. Middleware layer is an approach to satisfy the need of wireless sensor network at the level of design and implementation. Various approaches have been proposed as a middleware in wireless sensor network [8]. These middleware architecture deals with dissimilar applications like installation of software in efficient manner and aggregation of data. Sometimes application needs more advanced functions to reduce the complicated design and implementation with the middleware architecture. In the current dynamic scenario Service oriented architecture is an approach which may attempt these challenges. This approach provides a flexibility which is very much useful to add a component or replace the old component to fulfil the demands of application. Various services [9] can be managed at network level with wireless sensor network by way of service oriented model i.e. Flexible Service Oriented Model. Very few researchers have done the survey on the middleware architecture for wireless sensor as well as network architecture [8]. Integration of heterogeneous platforms with a single standard interface is the demand of current era to support technological development [12].

SOM is a possible way to do such kind of things because of its property. To perform the integration among various components flexibility should be there. So In this paper we have proposed a Flexible Service Oriented Network Architecture which is based on the approach of basic Service oriented architecture [19] to provide flexibility, loose coupling and reusability. Flexible Service Oriented Network Architecture consists of following components: 1. Building Blocks, 2. Building Block Repository, 3. Workflow Engine.

The proposed Flexible Service Oriented Network Architecture is able to couple or decouple the components according to the requirement of applications. In this paper we discuss the Service oriented architecture, issues of wireless sensor network with middleware architecture, Related works, Proposed Flexible Service Oriented Network Architecture for WSN as middleware, Comparison and Conclusion.

1.1 Service Oriented Model

In this model service provider play a role to develop new service module, Service broker registers the service with its description in UDDI (Universal Description Discovery and Integration), Service requester requests for service to service broker as per its need. Service oriented architecture is valuable to make a thing possible in heterogeneous environment. These advantages are: It can provide a common interface to publish or subscribe the service in the heterogeneous environment; many applications can run on different platforms and interact easily with each other, Service user can be coupled for time duration according to need of application, It can reduce the dependency among the components due to loose coupling, Service user can consume an exact service without knowing its location.

From the above features of Service oriented model we can use those models which can work as middleware architecture. This type of middleware is capable to integrate the components

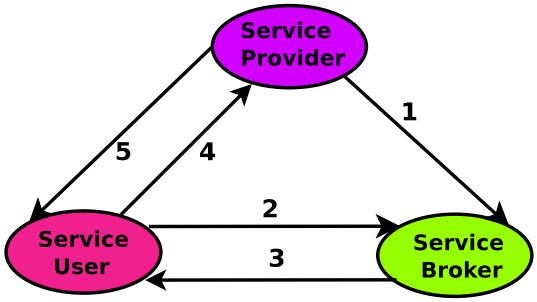


Figure 1: Service Oriented Model

and components reusability on demand in heterogeneous computing environment. It can handle the functional (service creation, communication, invocation, group management, and real-time operation) and non functional (service discovery, interoperability, reusability) requirements of application. Numerous middleware architectures have been proposed to add a new functionality like enterprise system. This approach is used to provide reusability of components like protocol, software etc. Service Oriented Model based middleware is a demand of todays environment.

2 Issues of Wireless Sensor Networks with Middleware Architecture

Middleware architecture is software infrastructure which will provide connectivity between components like hardware, operating system, communication protocol, network corresponding to application. Middleware architecture should be capable to provide runtime support, connectivity and execution of applications in dynamic environment. Middleware architecture provides isolation among various components at abstraction layer. Various challenges are present to develop service oriented model based middleware architecture for wireless sensor network. Challenges are consist of two things: one is development of middleware and another is wireless sensor network. Development of Flexible Service Oriented Network Architecture is a challenging issue which will support the applications of wireless sensor network. Some issues are discussed below.

Service Heterogeneity: Heterogeneity is a major problem in wireless sensor network at application level. This heterogeneity can be defined in terms of variation of services, platforms and networks. Services heterogeneity is increasing to fulfill the need of users requirements. Various services are present to perform the same operation in wireless sensor network at different levels of application. It is very much difficult to provide standard service in the service oriented environment to achieve the quality of service. Therefore Service oriented environment needs a common architectural platform for wireless sensor networks, which can provide a method to access the services form different platforms.

Adoption of Protocol: Requirement of users are increasing according to application, consequently number of protocols are also increasing to full fill the requirements. In current environment users or programmers are developing a protocol to support the application requirements. Number of protocol will present on the web in future but protocols may not be common to satisfy the requirements of application. Maintenance of protocol is also an issue. To address these issues, flexible middleware architecture is needed. Flexible middleware architecture should provide loose coupling. Loose coupling will provide flexibility to couple or decouple the protocols with application requirement. So flexible middleware architecture will also helpful for maintenance of protocol and users will have flexibility to choose the protocol from the web according to need.

Integration with Networks: Integration mechanism will provide interaction to share the services or information with other network according to requirements of application. So middleware architecture should have capability to integrate the WSN application with external environment. Middleware architecture can provide interface for communication.

Interoperability: In Service oriented environment service vendors, service providers and platforms are increasing to perform the task in Wireless sensor network. These services are developed on different platforms such as C, java etc. Interoperability between different platforms should be present to provide a service composition. Therefore Wireless sensor network needs interoperability mechanism to support different type of services and platforms.

Adoption of New Service: Innovation of new technique and improvement of previous technology provides new services for various application domains such as Wireless sensor networks, Adhoc networks, Cloud computing etc. These developments are the result of requirements and feedback of users after using the applications. All platforms may not be support the newly developed services. Therefore Adoption of new service is one of the crucial issues in Wireless sensor network.

Security: Sensor network is use to capture the data or information from the environment. Information or data can be sensitive according to application like military services, health services etc. This type of application requires security modules to secure the data from attacker. So security is also an issue to provide secure communication and data between WSN and system. Service oriented model can provide security modules as a service as per requirement. Very few works have been done on the security with service oriented model [2]. Therefore Flexible Service Oriented Network Architecture will suit to resolve the security issue.

3 Related Works

Various methods have been proposed to address the issues of wireless sensor network. This section provides overview of various solutions in the context of wireless sensor network issues and its solutions. These solutions are serving the way to tackle the problems. Wireless sensor network are suffering from the problem of data aggregation, resource management, energy efficiency, heterogeneity, hardware, routing etc. To solve these issues authors have been proposed various approach in terms of middleware technology. Middleware approach provides abstraction at various levels such as interoperability, language, hardware, complexity etc. We are discussing the some work in comprehensive mode.

COMiS: This is component based middleware architecture [10] which is used to satisfy the constraints like power and memory. Components of this middle- ware can be varying as per the behaviour of sensor nodes and loaded memory based application. Components of this middleware are Listener, Register, Discovery, Send and update. Each component is responsible to perform its task according to functionality. It provides a various additional functionality such as discovery of k components according to distance, component management, and registration and component updation. It is more suitable for collaborative applications.

SOMA: An SOMA [1] tries to fulfill the gap of interoperability among homogeneous platform. This middleware (Service Oriented Middleware Architecture) approach provides a communication model for wireless sensor network. To reduce the power consumption this solution is feasible at high traffic. Standards of middleware components are inspired with the web services, which will helpful to accommodate the component interoperability and reusability. This solution provides a way to manage the homogeneous platform with interoperable manner. It has some advantages like complexity reduction, service specific interface etc. In this approach major drawback is gateway dependency because failure of single point will affect the performance of whole network.

MiSense: It is Component based middleware layer [13] which is use to support distributed sensor application. MiSense reduces the complexity of system by imposing structure on top of component model. This middleware provides a resource management, network abstraction and communication without knowing the detail of low level complexities. It is a feasible to handle the issues like data aggregation, topology management and event detection.

SOM: Service oriented model [3] is much efficient than traditional middleware. Service oriented computing environment provides better availability, easy accessibility with some standard models and protocols. This approach provides a facility of interoperability and loose coupling between services in distributed scenario. SOM is capable to handle the communication, service discovery and publication among various services.

OASiS: OASiS is a programming framework [14] which provides abstraction layer to hide the complexity at low level. This is planned for resource constraint application devices. OASiS framework is combination of various service modules such as Composer, Node manager, Object manager, Service discovery protocol. These modules are responsible to perform the specific task. Node manager is responsible for routing, service discovery is responsible for search the service domain, Object manger is use identify the object and Composer is responsible for initiating a service graph, parsing of graph and binding of service. This framework is feasible to support the single object at particular instance but it can be improve.

SMC: This framework is based on service oriented architecture [4]. In this middleware various components are defined as a service. These components are use full to support various application domains. Service based midtier component (SMC) is capable to deal with different functions. This model consists of layers and each layer is responsible to perform particular task. Layers are independent to each other. Service bus layer provides a facility to establish communication between protocols. Binding is responsible to bind the service according to application. Interface repository service layer is used for repository and classification of domain. Data exchange service provides facility to exchange the data among similar nodes. Encryption is used for security of data. Synthesizer service layer is use to facilitate the processing of data and its composition. Routing layer is responsible to handle various queries at same time.

USEME: USEME is programming framework [7] which is based on service oriented approach. This framework provides facility to developer to deploy the service on sensor node and actuator. These services are developed to satisfy the real time constraint and specification. It is platform independent. This framework support to wireless sensor network as a middleware. Useme framework is combination of various components. Constraint management is use full to fulfill the demand of real time environment, Configuration management is responsible to maintain the frequency of service discovery, Invocation and communication is used to deliver the packet among various nodes, Publication and service discovery is used to store the record of service and group management is responsible to form a group to achieve the efficiency and scalability.

GEM: GEM is a generic event service middleware framework [11]. This framework is designed to support the new event based service. Gem is generic middleware, which is use to facilitate the wireless sensor application along with service package. Generic middleware architecture provides multilevel event detection, event language description and mote module to encode the code. Language of event defines the format of data and storage in mote. Information of every mote is required to identify the event at group level.

A range of solutions are existing to address above issues.These solutions are capable to resolve the above issues of wireless sensor network.Some issues of wireless sensor networks still exist such as adoption of new services, new protocols, and interoperability between heterogeneous services.Therefore we are providing a Flexible Service Oriented Network Architecture as a solution which is capable to tackle these issues.

4 Proposed Work

This is generic architecture of the system with proposed Flexible Service Oriented Network Architecture. Generic architecture consists with four layers (Figure 2). Every layer is responsible to perform specific task as per the need of application. Proposed Flexible Service Oriented Network architecture exists on top of every layer. Other layers are application software, X- operating system and system software. Application s/w is repository of softwares like web browsers, database spread sheet etc, which provides a communication between proposed architecture and X- operating system. X- Operating system is component based operating system. System s/w provides a communication between h/w and X- Operating system.

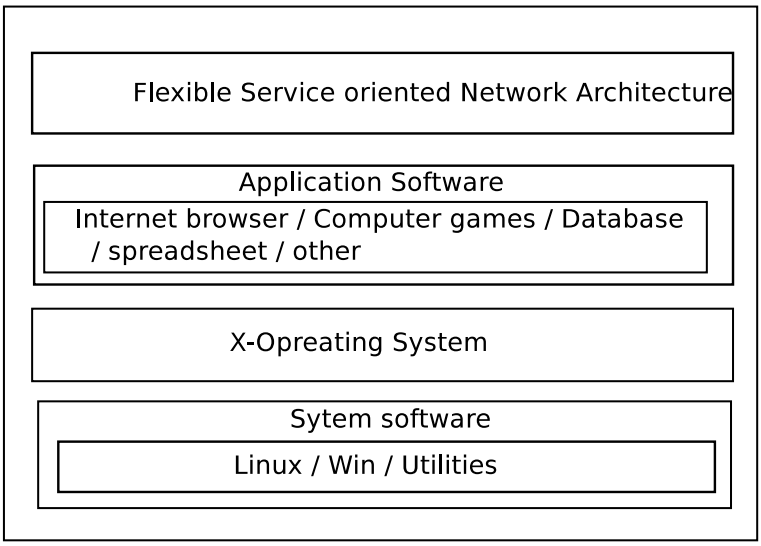


Figure 2: Generic System Architecture

In proposed Flexible Service Oriented Network Architecture, a new solution for the Wireless sensor network which is constraint based and constraint free. Constraint based and constraint free network can have various issues such as adoption of new service, adoption of new protocol, Interoperability, service selection and composition availability of service, security, heterogeneity of platforms. Detail functionality of Flexible Service Oriented Network Architecture in Figure 3.

Service user will be able to find out the service as per the requirements of application. Service user can be classified in terms of domain expert and general user with its knowledge. Users can give input to API with requirement such as application requirements, user requirement, and network requirements.

Service broker will get the request through the API. Broker will search the service in its local repository as per the requirement. If service will available then broker will return optimized service for application otherwise send the request to another broker to get best service. Service provider will provide the service to user through broker. Service provider will compose the service as per the user requirements and reply back to broker. Broker will send the service to user without knowing the low level detail of service.

Completion of above steps user will get the best service as per the requirements. Then user will be able to deploy the service for wireless application like sensor network through the API.

Service user: In this architecture level of users are classified in two categories such as Non parametric user and parametric user based on expertise. The execution of services in flexible

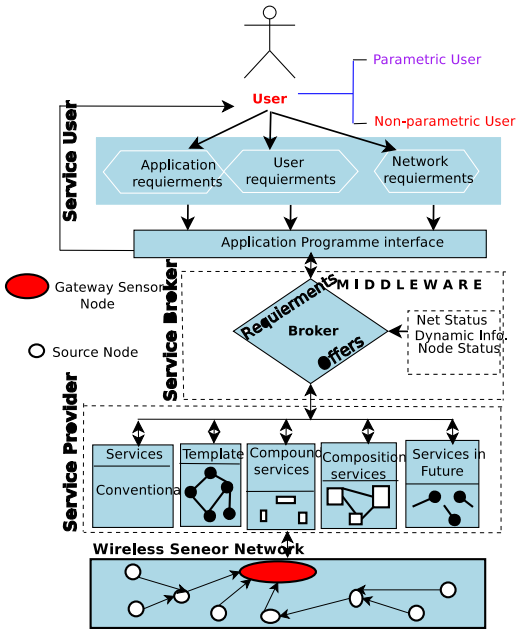


Figure 3: Flexible Service Oriented Network Architecture

service oriented network architecture begins with requirements of application as a user input.

Non Parametric User: Non parametric users are passive recipient of service without knowing the technical or low level details. The service request of Non parametric user is forwarded to broker through the application programming interface without any technical requirements; the requested services are fetched from service provider with default setting and delivered to the user through the broker. In this case non parametric user may not get filtered services due to lack of technical requirements as an input.

Parametric User: Parametric users are service users, who have detailed technical knowledge of application requirements. These requirements are application requirements, user requirements and network requirements. A parametric user sends request in the form of requirements parameters of application to the service broker. Service Broker will fetch the service from the service provider after composition as per the requirements and reply back to the Service users. Composed service will be filtered and most suitable for the application. Service users will deploy the service in application after getting the reply from broker.

Service Broker: There three types of functionalities, which are performed by a broker to complete the success full communication in flexible service oriented environment.

Protocol Graph Generator: A specific pattern of combination or sequence of building block known as protocol graph. Complex services can be made with the combination of various simple service s. A graph sketch specifies the specific pattern, which is needed for getting a service of user. Every node has capacity to generate the protocol graph. Requirements like user, application, and network are playing important role for dynamic generation of protocol graph.

Message Translator: In this module some messages are composed with the combination of more than one message, which is performed by the message list. A building block can add message, create message read message and removes the message from message list as per the compatibility. Application building block is used for message transformation. This building block also plays a role of bridge between workflow and application.

Message Transmission: This module is responsible for the transmission between application and Network building blocks. Message Formats for communication Following pattern is

used for communication in flexible service oriented network environment.

msg 1 and msg2: A XML file name in string format.

msg4 and msg5: This is msg1,built in message list format.

msg3 and msg6: Tagged message with filename and fragmented data to be transferred.

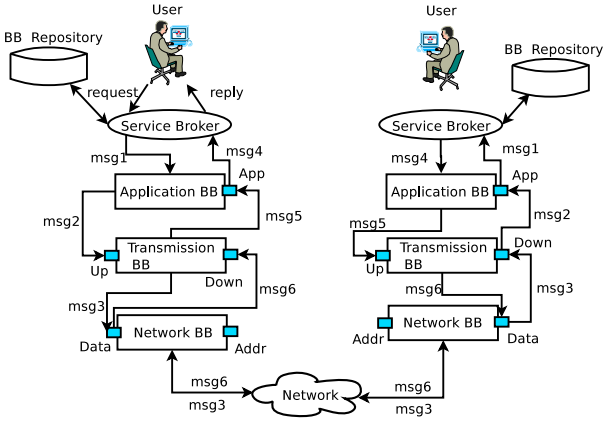


Figure 4: Communication Process between Brokers

After receiving the request from user end, Broker generates a list of unavailable building blocks (Service module) in XML format. Building blocks information sent to application building block in the form of msg1 (Figure 4) [15]. Up port of Transmission building block receives a msg2 from application building block after translation. Transmission building block forwards msg3 to the data port of Network building block. Then received msg3 sent to the other Network Building block with same msg format. At the receiver end, Network building block sends msg to the down port of transmission building block. The data pull out from Transmission building block, will be copied in XML format. Above process will be repeated for reply msg in appropriate format.

Service Providers: The service providers are entity of our proposed architecture, which provides service to user as per the requirements. These services are classified according to its nature in various categories such as Compound, composition, conventional, template and services in future. A Service provider plays a role of developer and provider for consumer. Service provider provides a many flavour of service along with same or different providers. Service providers have facility to develop a new service based on predefine, precompiled and precomposed methods as per the requirement parameters of users. These above methods are helpful at the time of service composition. The service provider can use static or dynamic methods to develop a new service. When parameters of service need not change frequently at run time then developer follows the static method otherwise dynamic method to support the frequent change at runtime.

Services: Service is functional module, which is used to perform single or multiple tasks. Functional module of service may be a single or combination of other modules. These modules can be implemented by any technology on any platform with scalability, flexibility and fault tolerance. The goal of concerning things are interoperability, heterogeneous system services should be capability to connect, loose coupling, to minimize interdependencies between modules and heterogeneous data type, interaction pattern at the time of platform binding, logical deployment. These are performed with loose coupling. Our Flexible service Oriented Network architecture supports for wide range of wireless sensor network services.Conventional Service Represents those type of services which developed in previous days and we are using without any required modification. These services have some flaws like network protocol adaption, Localization algorithms and its complexity, tight coupling etc.

Template: Template is a service, which is defined as set of reusable service related data that will minimize the required time to create service. This is applied for designer, developer and user to create or use the service. The template is used by various applications in order to save time needed to develop network interface, network load balancer and one or more virtual machines. This architecture provides a facility for developer, architect to develop or design new template as a service and register it for use of public environment.

Compound Service: The Compound service is a result from combination of two or more modules, which may be different or same according to configuration. Development of compound service is not easy task because before the development of this service user should know the compatibility of application programming interface, protocol and integration requirements for application. Prior knowledge of low level detail of different services and platform is not easy. So there is need of architecture to support in the development and deployment of compound services for application. Our Architecture resolves above issues and provides support to developer and user without knowing the low level details of other services.

Service Composition: Service composition is a way to build a new service from a set of service modules at run time. Service location and selection plays an important role in the process of composition. In this architecture we are considering two types of service compositions. Static service composition all requirements are fixed at the time of service designing. The requirements cannot be changed frequently at run time because lack of flexibility, agile for run time. Therefore we are also considering dynamic service composition to accomplish the requirements of user as a service. In dynamic service composition various complications comes like operating service, service selection criteria. It is not necessary that composition process will always give correct result as per the user requirements. So through the application programming interface provides a facility to change in the requirements to get an available service. Dynamic service composition reduces the human effort in service selection. We are considering both environments for service selection in our proposed architecture. Service in future Services in future are vision to develop a new service, protocol, tools, plat- form, language etc. In the current era of development researchers and developers are providing a new solution for the specific problem as a service to fulfill the demand of users with fast speed. So services are increasing with rapid rate on the internet. Users may not know all solutions of specific problem. Therefore we are providing architecture which will give the best service which will available on internet in future.

Technical Requirements: Many solutions are available in the form of service for specific task on the web. These services are required to execute the application. User will have flexibility to give an input in terms of requirement for service selection and composition [18]. Therefore user has option to select a service with specification of requirements. Being a user, requirements can be categorized such as user requirement, application requirement and network requirement which are given in Figure 5.

User requirements can be defined as a goal of application.This requirement is useful to find out the type of service for application [17]. There are various type of applications are used in sensor network so requirements of application will also be differ as per the performance. Application requirements may be defined as accuracy, computation time, bandwidth, delay etc. network requirements is defined as, need of network related constraints. These constraints may be the selection of protocol, level of fault tolerance, transmission speed, life time of network etc. These requirements play an important role to find out the best service for the application.

Let S be the set of services which is represented as:

$$S=\sum_{i=0}^{l} U_i + \sum_{j=0}^{n} A_j + \sum_{k=0}^{m} K_m.$$

User requirement may be single or multiple, which is mandatory. A is application require-

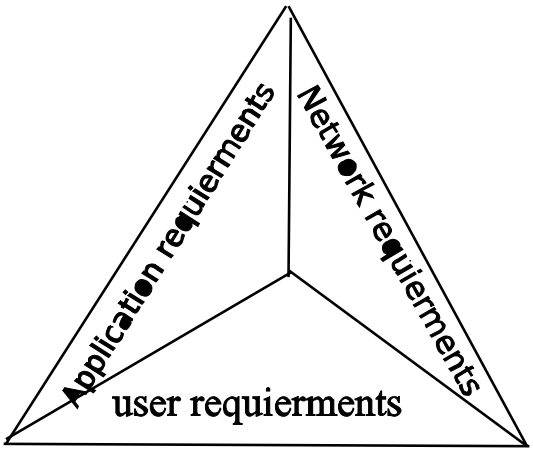


Figure 5: Technical Requirements

ments, N is network requirements.These requirements are optional, which can be represented as:

$$A=\{a_1, a_2, a_3, ...a_n\}$$
$$N=\{n_1, n_2, n_3, ...n_k\}.$$

When n, k=0 then interaction pattern of application is similar to current pattern, where user will get minimal facility of service selection. When j=1...n and k= 1...m, then this condition will increase the better tune of service. The application uses different protocols for communication and also there are many applications which can work with diverse of protocols.

Let P be the total number of protocols available which can be represented as:

$$P=\{p_1, p_2, p_3.....p_n\}\ ,\ a_f=\{p_f | p_f \subset p\}.$$

Convergence of service request interpretation:

$$E(X-S)^2=0$$

then X='S' with probability 1 or x converges to S with probability 1 let $\sigma^2 x$ is variability of measurement according to Chebyshev inequality:

$$\mathrm{p}\{|X-A| \leq \epsilon\} \geq |\sigma^2 x \div \epsilon^2 \mathrm{x}.$$

If σ_x is very much smaller than ϵ then observed variable X is between (S-ϵ) and (S+ϵ) which is almost certain and one measurement of X is sufficient. However if σ_x is not sufficiently smaller compared to ϵ, then the result will not be of sufficient accuracy.

5 Comparison of Various Service Oriented Approaches with Flexible Service Oriented Architecture

Wireless Sensor network is distributed service over the network with well-defined interface. These services are easily accessible and deployable by WSN application through service oriented model based Interfaces. Service oriented model covers functional and non functional requirements of users to fulfill the gap of applications. These applications can be use the services to get the better result as per the requirements. To achieve result communication should be perform in between two parties with indirect manner i.e. with service brokers. Services are accessible through multiple paths; best path will increase the response time of service with UDP, SOAP transport. In the comparison table 1. We have tried cover all the requirements of user and applications with Flexible Service Oriented Architecture for WSN. Few of them are covered in individual model but no one covers all the requirements and flexibility of module implementation

as per the new requirements, which may come in future. Therefore we have shown comparative analysis in Table 1.

Table 1: Comparison with main features

SI. No.	SOA based Approach	Environment	Features	Covered Requierments
1	Flexible Service Oriented Network Architecture for WSN	Wireless/Wires Sensor Networks, Supports TCP/IP based networks	Flexibility to Innovation, Loose Coupling, Automatic service discovery, Adaption of New service, Support real time constrains of users	Run time support,Service transparancy, Interoperability among service,QOS,Integration with other sytem,Service discovery, Service abstraction.
2	USEME	Wireless/Wires Sensor networks	Handles publicationand discovery,Management of real time constraint	Runtime support service discovery Service abstraction
3	OASiS	Wireless Sensor Networks	Environment based sepration Concerns,Dynamic service discovery and deployment	Runtime support system, Service discovery,Abstraction.
4	Misense	Wireless Sensor Networks	Content based on sepration of concers and subscription model, Flexible data management,Programming API	Service transparancy,Efiicient large amount of data.
5	Stream Ware	Sensor network (Wired/Wireless)	Query based access. Hetroginity Management,Scalability	Service discovery,management,Efficient handling of large amount of data.
6	(SI) 2	Smart items Network	Platform independent service, description, deployment	Run time support, service discovery and deployment. Abstraction, Integration with other system.
7	DySSco	Wireless sensor network	Dynamic self configuration service coverage	Service deployment, transparency, abstraction, Configurable service.
8	B-VIS	Distributed RFID and Sensor networks	Real time Tracking and monitoring trough sensors, Programming for real time data.	Abstraction ,Interoperability, Efficient for handling large amount of data. Integration with other system.
9	SOMDM	Wireless Sensor Network	Component architecture to reduce data processing	Service transparency, abstraction to heterogeneous nodes, interoperability between other nodes.

6 Conclusions and Future Works

Our Proposed Flexible Service Oriented Network Architecture is developed with java platform. XML format is used to provide interoperability between different services. Flexible Service Oriented Network Architecture is providing a facility to user for service selection and composition, adoption of new service, protocol. Through this architecture users can get best service from the broker as per the requirements of application.

Also, the architecture provides a facility to couple the module (service) in the real time service domain and feasible to connect the heterogeneous and homogeneous services on common

platform. This architecture can be used by novice users having no or little knowledge regarding the low level technical detail of services. In future we will convert the XML format to RDF format, rating of service on constraints.

Bibliography

[1] Abangar H. et al. (2010); A service oriented middleware architecture for wireless sensor networks, *Proc. of future network and mobile summit conference*, 1-10.

[2] Al-Jaroodi J., Al-Dhaheri A. (2011); Security issues of service-oriented middleware, *International Journal of Computer Science and Network Security*, 11(1): 153-160.

[3] Al-Jaroodi J., Mohamed N. (2012); Service-oriented middleware: A survey, *Journal of Network and Computer Applications*, 35(1): 211-220.

[4] Alkazemi B.Y. (2012); Support Cross-domain Components Composition in Software Production Lines, *7th International Conference on Software Paradigm Trends (ICSOFT 2012)*, 421-426.

[5] Alkhatib A. (2012); Wireless Sensor Network Architecture, *International Conference on Computer networks and Communication Systems (ICNCS 2012)*, 35: 11-15.

[6] Dai L., Xu H.K., Chen T., Qian C., Xie L.J. (2014); A Multi-objective Optimization Algorithm of Task Scheduling in WSN, *International Journal of Computers Communications & Control*, ISSN 1841-9836, 9(2): 160-171.

[7] Canete E., Chen J., Diaz M., Llopis L., Rubio B. (2008); USEME: A Service-oriented Framework for Wireless Sensor Networks and Actor Networks, *Proc. ASWN2008*, 47-53.

[8] Hadim S., Mohamed N. (2006); Middleware challenges and approaches for wireless sensor Networks, *IEEE Distributed System Online*, 7(3): 1-23.

[9] Han G., Xu H., Duong T.Q. (2011); Localization algorithms of Wireless SensorNetworks: a survey, *Telecommunication system*, Springer, 52: 2419-2436.

[10] Janakiram D., R. Venkateswarlu R. (2005); A Distributed Compositional Language for Wireless Sensor Networks, *Proc. of IEEE Conference on Enabling Technologies for Smart Appliances (ETSA)*, 1-11.

[11] Jiao B., Son S.H., Stankovic J.A. (2005); GEM: Generic Event Service Middleware for Wireless Sensor Networks, Available online `http://www.cs.virginia.edu/~stankovic/psfiles/binjia-inss.pdf`, 1-6.

[12] Lozano-Garzon C., Ortiz-Gonzalez N., Donoso Y. (2013); A Proactive VHD Algorithm in Heterogeneous Wireless Networks for Critical Services, *International Journal of Computers Communications & Control*, ISSN 1841-9836, 8(3): 425-431.

[13] Khedo K.K., Subramanian R.K. (2009); A Service-Oriented Component-Based Middleware Architecture for Wireless Sensor Networks, IJCSNS *International Journal of Computer Science and Network Security*, 9(3): 174-182.

[14] Kushwaha M., Amundson I., Koutsoukos X., Neema S., Sztipanovits J.(2007); OASiS: *A programming framework for service-oriented sensor networks*, Technical Report, USA.

[15] Manu A.P., Rudra B., Kumar V., Vyas O.P. (2012); Broker's Communication for Service Oriented Network Architecture, *International Journal of Future Generation Communication and Networking*, 5(4): 77-88.

[16] Marks M., Niewiadomska-Szynkiewicz E. (2009); Multi objective approach to localization in wireless sensor networks, *Journal of Telecommun. Inform. Technology*, 3: 59-67.

[17] Niewiadomska-Szynkiewicz E.,Marks M. (2009); Optimization schemes for wireless sensor network localization, *Appl. Math. ComputerScience* 19: 291-302.

[18] Oppermann F., Peter S. (2010); Inferring technical constraints of a wireless sensor network application from end-user requirements, *Mobile Ad- hoc and Sensor Networks (MSN), Sixth International Conference on IEEE*, 169-175.

[19] Reuther B., Henrici D. (2008); A model for service-oriented communication systems, *Journal of Systems Architecture*, 54(6): 594-606.

[20] Sifalakis M., Louca A., Bouabene G., Fry M., Mauthe A., Hutchison D. (2011); Functional composition in future networks, *Computer Networks*, Special Issue on Architectures and Protocols for the Future Internet, 55(4): 987-998.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 9(5):623-632, October, 2014.

Steganalysis of $\pm k$ Steganography based on Noncausal Linear Predictor

K.M. Singh, Y.J. Chanu, T. Tuithung

K. Manglem Singh
Department of Computer Science & Engineering,
NIT Manipur, Imphal, India
Email: manglem@gmail.com

Yambem Jina Chanu*, Themrichon Tuithung
Department of Computer Science & Engineering,
NERIST, Itanagar, India
*Corresponding author: jina.yambem@gmail.com
tth@nerist.ac.in

Abstract: The paper proposes a novel steganalytic technique for $\pm k$ steganography based on noncausal linear predictor using prediction coefficients obtained from the autocorrelation matrix for a block of pixels in the stego-image. The image is divided into equal-size blocks, autocorrelation matrix is found for the block, and the appropriate noncausal linear prediction coefficients is selected to predict all pixels in that block. A pixel is assumed to be embedded with message bit if the absolute difference between the original pixel value and predicted pixel value exceeds the pre-defined threshold. The effectiveness of the proposed technique is verified using different images.
Keywords: LSB embedding, noncausal linear predictor, RS analysis, steganalysis, steganography.

1 Introduction

Our world is becoming smaller and smaller day by day due to the rapid development in computer and network technology, empowered with unlimited expansion of communication and information technology capability globally for exchange of information with large involvement of individuals at home, private and public sectors and government organizations. Both business and society now are highly dependent on the Internet and multimedia technology as an integral part for communication, which on other hand serves two purposes - one for the exchange of vast amount of information and knowledge for the welfare of human beings, and another for the destruction of humanity by using such medium. Law enforcement personnel and criminals conceal their message, work plan, and important information in digital media. This technique is called steganography, which is the art and science of embedding secret messages inside different cover media such as text, audio, image and video without any suspicion. The Internet itself provides hundreds of freeware and shareware, which can be freely used by anyone including criminals and terrorists. To detect and extract the hidden message from stego-data by the cyber security personnel and law enforcement professional is the current interest of steganalysis, which is the science of discovery of existence of hidden information.

The race between steganography and steganalysis is ongoing and never-ending competition, in which the former develops newer and more robust algorithms to embed secret message into cover medium such as text, audio, image and video to form stego-data, which is not different from the original medium perceptually to the ordinary person, and does not arouse any suspicion, while the later tries to stop all steganographic techniques or detect the embedded secret message from the stego-data. Two important media which appeal to hide secret message are the image and video due to the inherent presence of high redundancy in representation of such

data. The most obvious of steganography is to hide data, but its general applications include covert and invisible communication between two or more parties, secret data storing in storage devices, integrity checking of data by using hash value, access control mechanism for digital content distribution by including access key for selective extraction, and media database system by unifying the media data such as music, picture, video with other information like date and time of recording such data as metadata [1].

Steganalysis finds applications in gathering and tracking criminal and cyber terrorist activities and anti-social elements over the Internet, cyber forensics investigation for extraction of hidden message from compromised digital systems, cyber warfare for waging war using the Internet, peaceful purposes for improving steganographic tools to identify their weakness [2]. In the recent decades, cyber security community and professionals take challenging, interesting and exciting researches in developing new and robust steganographic algorithms as well as better steganalytic algorithms to counter the steganography [3–5]. The powerful and popular LSB detection algorithms are Chi-square [6], RS [7], Gradient Energy-Flipping Rate (GEFR) Detection [8] and Histogram difference [9]. These algorithms can estimate the length of the embedded message in LSB embedding. LSB matching (LSBM) steganographic technique adds or subtract by 1 if LSB does not match with message bit [10] [11]. LSBM is a special case of k steganography with *k=1* [12]. In k steganography, pixel values are either increased or decreased by k. k steganography is an important steganographic technique, and number of steganalytic algorithms for k steganography particularly based on linear prediction is very few. This motivates to develop this novel k steganalytic algorithm.

This paper proposes a new steganalytic technique for k steganography using noncausal linear predictor based on prediction coefficients obtained from the autocorrelation matrix for each block of pixels in the image. Most appropriate elements from that matrix are used for predicting all pixel values in that block. The difference between the predicted pixel value and the original pixel value gives the knowledge about the presence of message bit.

The paper is organized as follows. Section 2 deals with a short introduction on the order of prediction, which is used in the proposed algorithm. Section 3 proposes the new steganalytic algorithm, followed experimental results in Section 4 and conclusions in Section 5.

2 Order of Linear Prediction

The order of linear prediction in images can be based on the neighbourhood set G, which is defined at a coordinate *(m,n)*$\in G$ with the points lying within a specified radius as its elements [13]. An η^{th} order neighbourhood set at coordinate *(m,n)* is defined using the Euclidian distance for defining the radius as

$$S^{\eta}_{(m,n)} = \{(i,j) : 0 < (m-i)^2 + (n-j)^2 \leq D_\eta\} \tag{1}$$

where *(i,j)* is the coordinates of pixels in the neighbourhood set, $(i,j) \neq (m,n), D_\eta$ and is an increasing integer function of η. Take for examples

1) 1st order neighbourhood set : $\eta = 1; D_\eta = 1;$
   $S^1_{(m,n)} = \{(m-1,n),(m+1,n),(m,n-1),(m,n+1)\}$

2) 2nd order neighbourhood set : $\eta = 2; D_\eta = 2;$
   $S^2_{(m,n)} = \{(m-1,n),(m+1,n),(m,n-1),(m,n+1),(m-1,n-1),$
   $(m+1,n+1),(m-1,n+1),(m+1,n-1)\}$

Similarly, the 3rd and 4th order neighbourhood sets are defined with $\eta = 3$, $D_\eta = 4$ and $\eta = 4$, $D_\eta = 5$ respectively. The neighbourhood sets up to order 5 for linear prediction is shown in Figure. 1. The pixel to be predicted is labeled '0'. For the first-order prediction, the neighbourhood set consists of pixels marked '1', for the second-order prediction, the neighbourhoods set involves pixels marked as '1' and '2' and so on. Figure. 1 shows the causal and noncausal regions of supports with pixels lying to the left of dark line for the causal prediction.

3 Proposed Steganalytic Algorithm

Linear prediction is used for modelling, estimating and coding of one-dimensional signals, in speech coding and understanding, geophysics and biomedical signal processing applications [14], and of two-dimensional signals in image compression, image segmentation and classification and spectrum estimation [15]. Linear prediction model is used in estimating pixels at any location of the image due to strong correlation among neighbouring pixels in the image. Most of the applications of linear prediction use the causal linear prediction, which predicts the current pixel value based on its surrounding past pixels only. Sometime, same prediction coefficients are used to predict entire pixels for the same block in the image.Asif and Moura wrote that the noncausal linear prediction, which is based on both the past and future pixels in the neighbourhood of the current pixels give better results [13].

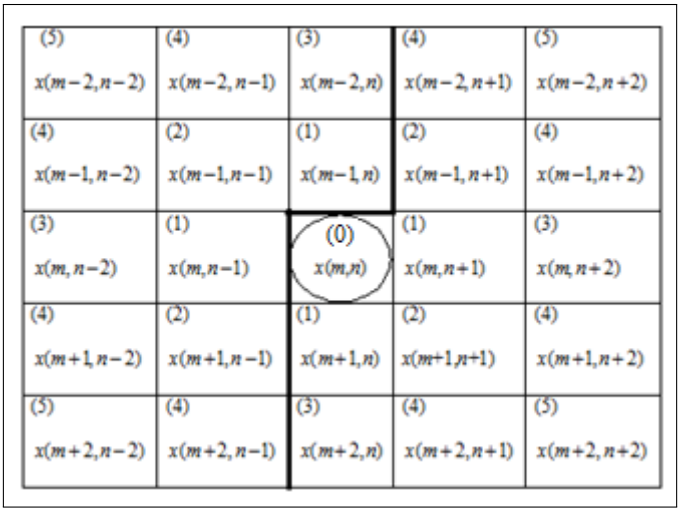


Figure 1: The neighbourhood sets for the 1^{st}-order to 5^{th}-order prediction.

The predicted pixel $\hat{x}(m,n)$ at the location (m,n) is given by

$$\hat{x}(m,n) = \sum_{(i,j)\in W} \sum a(i,j)x(m-i,n-j) = x_\eta a_\eta \tag{2}$$

where W represents the prediction window that excludes $(0,0)$ and $a(i,j)$ is the prediction coefficient, x_η is a row vector of pixels used for prediction and a_η is the column vector formed by the corresponding prediction coefficients. The linear prediction error *(LPE)* is given by

$$e(m,n) = x(m,n) - \hat{x}(m,n) = x(m,n) - x_\eta a_\eta \tag{3}$$

The corresponding *mean-square error* $E(e^2(m,n))$ is given by

$$E(e^2(m,n)) = E(x(m,n) - x_\eta a_\eta)^2 = E(x^2(m,n)) - 2\acute{a}_\eta E(\acute{x}_\eta x(m,n)) + \acute{a}_\eta E(\acute{a}_\eta x_\eta)a_\eta$$

$$= R_x(0,0) - 2\acute{a}_\eta r_{x_\eta} + \acute{a}_\eta R_{x_\eta} a_\eta \tag{4}$$

where $R_{x_\eta} = E(\acute{x}_\eta x_\eta)$ is the autocorrelation matrix, $r_{x_\eta} = E(x(m,n)\acute{x}_\eta)$ is its cross-correlation vector with the pixel to be predicted and $R_x(0,0) = E(x^2(m,n))$. The optimal prediction coefficient vector is obtained by minimizing $E(e^2(m,n))$ with respect to a_η and found by the condition

$$\nabla(E(e^2(m,n))) = 0 \tag{5}$$

where ∇ is the gradient operator with respect to the vector a_η This yields in the following matrix form of *normal equations* or *Wiener* equations

$$R_{x_\eta} a_\eta = r_{x_\eta} \tag{6}$$

The solution of the above matrix equation gives the prediction coefficient vector a_η. Using the symmetry property $R_x(i,j) = R_x(-i,-j)$of the autocorrelation function of wide-sense stationary image random field, it can be shown that [16]

$$a(-i,-j) = a(i,j) \tag{7}$$

The symmetry property simplifies the computation of the prediction coefficients. The predicted pixel value $\hat{x}(m,n)$ should be different from the original pixel value $x(m,n)$ by k (where k is an integer value by which the value of the pixel at (m,n) is changed to hide a binary bit) if the message bit is hidden in the pixel value $x(m,n)$. The absolute error between the original and the predicted pixel values is given by

$$|e(m,n)| = |x(m,n) - \hat{x}(m,n)| \tag{8}$$

The total number of absolute errors, which exceeds a pre-defined threshold Θ is the estimate of length of hidden message bit in the stego-image. The following sub-sections are on different causal and noncausal linear predictors for the first and second order.

3.1 First-order Causal Linear Predictor

The predicted pixel $\hat{x}(m,n)$ of the centre pixel $x(m,n)$ for the first-order causal linear prediction is given by

$$\hat{x}(m,n) = a_1(0,1)x(m,n-1) + a_1(1,0)x(m-1,n) = x_1 a_1$$

$$= [x(m,n-1)x(m-1,n)] \begin{bmatrix} a_1(0,1) \\ a_1(1,0) \end{bmatrix} \tag{9}$$

The pixels $x(m,n-1)$ and $x(m-1,n)$, marked as "1" are shown to the left side of dark line in Figure.1, and used in the first-order causal linear prediction. The prediction coefficients $a_1(0,1)$ and $a_1(1,0)$ for the first-order causal predictor are related with the autocorrelation functions by the following normal equation in Eq. 6: $R_{x_1} a_1 = r_{x_1}$ where

$$R_{x_1} = \begin{bmatrix} R_x(0,0) & R_x(1,-1) \\ R_x(1,-1) & R_x(0,0) \end{bmatrix} and \quad r_{x_1} = \begin{bmatrix} R_x(0,1) \\ R_x(1,0) \end{bmatrix}$$

The number of different elements from the autocorrelation matrix is 4 for the first-order causal predictor.

3.2 First-order Noncausal Linear Predictors

The predicted pixel$\hat{x}(m,n)$ of the centre pixel $x(m,n)$ for the first-order noncausal linear prediction is given by

$$\begin{aligned}\hat{x}(m,n) &= a_1(0,1)x(m,n-1) + a_1(0,-1)x(m,n+1) + a_1(1,0)x(m-1,n) + a_1(-1,0)x(m+1,n)\\ &= a_1(0,1)x(m,n-1) + x(m,n+1) + a_1(1,0)x(m-1,n) + x(m+1,n) = x_1 a_1\end{aligned}$$

$$= \begin{bmatrix} x(m,n-1)+x(m,n+1) & x(m-1,n)+x(m+1,n)\end{bmatrix}' \begin{bmatrix} a_1(0,1) & a_1(1,0)\end{bmatrix}' \tag{10}$$

where $a_1(0,1) = a_1(0,-1)$, and $a_1(1,0) = a_1(-1,0)$ due to Eq. 7. The pixels $x(m,n-1), x(m,n+1), x(m-1,n)$ and $x(m+1,n)$ marked as "1" are used in the first-order noncausal linear prediction and shown in Figure. 1. The prediction coefficients $a_1(0,1)$ and $a_1(1,0)$ the first-order noncausal predictor are related with the autocorrelation functions by the following normal equation in Eq. 6: $R_{x_1} a_1 = r_{x_1}$ where

$$R_{x_1} = \begin{bmatrix} R_x(0,0)+R_x(0,2) & R_x(1,-1)+R_x(1,1) \\ R_x(1,-1)+R_x(1,1) & R_x(0,0)+R_x(2,0)\end{bmatrix} \textit{ and } \quad r_{x_1} = \begin{bmatrix} R_x(0,1) \\ R_x(1,0)\end{bmatrix}$$

The number of different elements from the autocorrelation matrix is 7 for the first-order noncausal predictor.

3.3 Second-order Causal Linear Predictor

The predicted pixel $\hat{x}(m,n)$ of the centre pixel $x(m,n)$ for the second-order causal linear prediction is given by

$$\hat{x}(m,n) = a_2(0,1)x(m,n-1) + a_2(0,1)x(m-1,n) + a_2(1,1)x(m-1,n-1) + a_2(1,-1)x(m+1,n-1)$$

$$= x_2 a_2 = \begin{bmatrix} x(m,n-1) \\ x(m-1,n) \\ x(m-1,n-1) \\ x(m+1,n-1)\end{bmatrix} \begin{bmatrix} a_2(0,1) \\ a_2(1,0) \\ a_2(1,1) \\ a_2(1,-1)\end{bmatrix} \tag{11}$$

The pixels $x(m,n-1), x(m-1,n), x(m-1,n-1)$ and $x(m+1,n-1)$, marked as "1" and "2" are shown to the left side of dark line in Figure.1, and used in the second-order causal linear prediction. The prediction coefficients $a_2(0,1), a_2(1,0), a_2(1,1)$ and $a_2(1,-1)$ for the second-order causal predictor are related with the autocorrelation functions by the following normal equation in Eq. 6: $R_{x_2} a_2 = r_{x_2}$ where

$$R_{x_2} = \begin{bmatrix} R_x(0,0) & R_x(1,-1) & R_x(1,0) & R_x(1,-2) \\ R_x(1,-1) & R_x(0,0) & R_x(0,1) & R_x(0,1) \\ R_x(1,0) & R_x(0,1) & R_x(0,0) & R_x(0,2) \\ R_x(1,-2) & R_x(0,1) & R_x(0,2) & R_x(0,0)\end{bmatrix} \textit{ and } \quad r_{x_2} = \begin{bmatrix} R_x(0,1) \\ R_x(1,0) \\ R_x(1,1) \\ R_x(1,-1)\end{bmatrix}$$

The number of different elements from the autocorrelation matrix is 7 for the second-order causal predictor.

3.4 Second-order Noncausal Linear Predictor

The predicted value $\hat{x}(m,n)$ of the centre pixel $x(m,n)$ for the second-order noncausal linear prediction is given by [17]

$$\begin{aligned}\hat{x}(m,n) = a_2(0,1)x(m,n-1) + a_2(0,-1)x(m,n+1) + a_2(1,0)x(m-1,n) + a_2(-1,0)x(m+1,n)+\\ a_2(1,1)x(m-1,n-1) + a_2(-1,-1)x(m+1,n+1) + a_2(1,-1)x(m-1,n+1) + a_2(-1,1)\\ x(m+1,n-1) = a_2(0,1)\{x(m,n-1) + x(m,n+1)\} + a_2(1,0)\{x(m-1,n) + x(m+1,n)\}+\\ a_2(1,1)\{x(m-1,n-1) + x(m+1,n+1)\} + a_2(1,-1)\{x(m-1,n+1) + x(m+1,n-1)\}\\ = x_2 a_2 \qquad (12)\end{aligned}$$

where $a_2(0,1) = a_2(0,-1), a_2(1,0) = a_2(-1,0), a_2(1,1) = a_2(-1,-1)$ and $a_2(1,-1) = a_2(-1,1)$ due to Eq. 7. The pixels $(m,n-1), x(m,n+1), x(m-1,n), x(m+1,n), x(m-1,n-1), x(m+1,n+1, x(m-1,n+1))$ and $x(m+1,n-1)$, marked as "1" and "2" are used in the second-order noncausal linear prediction and shown in Figure.1. The prediction coefficients $a_2(0,1), a_2(1,0), a_2(1,1)$ and $a_2(1,-1)$ for the second-order noncausal predictor are related with the autocorrelation functions by the following normal equation in Eq. 6: $R_{x_2} a_2 = r_{x_2}$ where

$$R_{x_2} = \begin{bmatrix} R_x(0,0) + R_X(0,2) & R_x(1,-1) + R_X(1,1) & R_x(1,0) + R_X(1,2) & R_x(1,-2) + R_X(1,0) \\ R_x(1,-1) + R_X(1,1) & R_x(0,0) + R_X(2,0) & R_x(0,1) + R_X(2,1) & R_x(0,1) + R_X(2,-1) \\ R_x(1,0) + R_X(1,2) & R_x(0,1) + R_X(2,1) & R_x(0,0) + R_X(2,2) & R_x(0,2) + R_X(2,0) \\ R_x(1,-2) + R_X(1,0) & R_x(0,1) + R_X(2,-1) & R_x(0,2) + R_X(2,0) & R_x(0,0) + R_X(2,-2) \end{bmatrix}$$

$$and \qquad r_{x_2} = \begin{bmatrix} R_x(0,1) & R_x(1,0) & R_x(1,1) & R_x(1,-1) \end{bmatrix}'$$

The number of different elements from the autocorrelation matrix is 13 for the second-order noncausal predictor. A block of the size $B \times B$ gives an autocorrelation matrix of the size $(2B-1)\times(2B-1)$. Selective elements for the autocorrelation matrix are given below. $R_X(0,0) = R_{X_2}(B,B), R_X(0,2) = R_{X_2}(B,B-2), R_X(1,-1) = R_{X_2}(B-1,B+1), R_X(1,1) = R_{X_2}(B-1,B-1), R_X(1,0) = R_{X_2}(B-1,B) R_X(2,0) = R_{X_2}(B-2,B), R_X(0,1) = R_{X_2}(B,B-1), R_X(1,-2) = R_{X_2}(B-1,B+2) R_X(1,2) = R_{X_2}(B-1,B-2), R_X(2,1) = R_{X_2}(B-2,B-1), R_X(2,-1) = R_{X_2}(B-2,B+1), R_X(2,2) = R_{X_2}(B-2,B-2), R_X(2,-2) = R_{X_2}(B-2,B+2)$ The first 7 elements in the above are used to predict all pixels in that blocks for the first-order noncausal prediction, and all elements in the above to predict all pixels in that block of the image for the second-order noncausal linear prediction.

4 Experimental Results

A set of 100 high-quality RGB images was collected from the Internet including satellite images, biomedical images, texture images, aerial images, photographic images, computer generated images and classic images commonly used in the color image processing literature. All images are of size 512×512 Figure. 2 shows representative images from this set. The performance of the proposed algorithm depends on the following various parameters and results are taken based on these parameters.

(I) Block-size of the prediction window (II) Noncausal or causal
(III) order of prediction (IV) Value of k

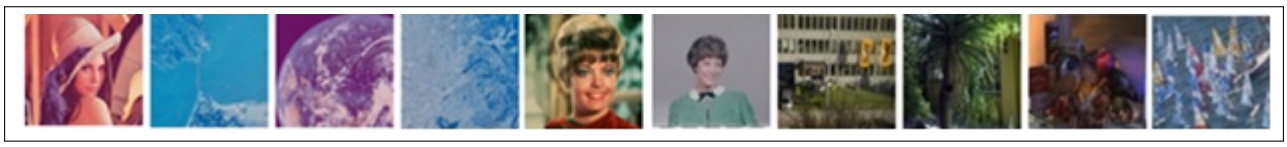


Figure 2: (a) Lena, (b) Golden Gate, (c) Earth, (d) Oakland, (e) Zelda, (f) Girl, (g) Building, (h) Aptus, (i) Hdr and (j) Sail.

4.1 Comparison of Performance based on Block-size

Experiments were taken to find the most appropriate block-size of the prediction using 100 different images. Results were taken for both the first-order causal and noncausal for the block-sizes of 4×4, 8×8, 16×16,32×32 and 64×64 respectively. These results in terms of percentage estimated length (EL) and percentage error (ER) are shown in Table1 and Table 2. Results were taken based on the averages of EL and ER of 100 images for *k=9*. The results in these tables are shown for the first-order. It was found that the performance is poor for both smaller and larger block-sizes. The most appropriate block-size is 16×16, which gives the least ER for majority of embedding percentages for both causal and noncausal predictors. The poor performance of the smaller block-size is due to the fact that the prediction coefficients obtained from smaller blocks can give accurate prediction if the blocks contain highly similar pixel values, but a block does not contain similar pixel values usually. On the other hand, larger block-sizes also give erroneous results, because the larger blocks contain many heterogeneous pixel values. Without loss of generality, block-size of 16×16 was used for all results afterward.

Table 1: Selection of the best block-size in terms of EL and ER for the first-order noncausal predictor

Block-size	Embedding Percentage										
	EL						ER				
	0%	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
4×4	-0.17	9.78	21.58	28.99	40.40	47.93	2.20	-15.80	10.10	-4.00	20.70
8×8	-0.64	9.77	22.83	28.88	39.56	50.58	2.30	-28.30	11.20	4.40	-5.80
16×16	-1.46	10.64	20.75	30.02	40.28	49.62	-6.40	-7.50	-0.20	-2.80	3.80
32×32	-0.94	10.13	23.50	29.69	40.28	46.60	-1.30	-35.00	3.10	-2.80	34.00
64×64	2.87	19.18	16.41	39.10	36.35	41.81	-91.80	35.90	-91.00	36.50	81.90

4.2 Comparison of Performance of Noncausal and Causal Predictors

Table 1 and Table 2 show that the noncausal predictors give less ER than causal predictor. The noncausal predictor uses all past and future pixels surrounding the pixels to be predicted and the causal predictor uses only the past pixels surrounding the pixel to be predicted. This results in better performance for the noncausal predictor.

4.3 Comparison of Performance based on Orders of Prediction

Table 3 gives the comparison of performance based on the orders of prediction. Results were based on the average results of all images for the causal and noncausal predictors with *k=9*. ER

Table 2: Selection of the best block-size in terms of EL and ER for the first-order causal predictor

Block-size	Embedding Percentage										
	EL						ER				
	0%	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
4×4	-2.51	12.53	23.03	28.63	38.95	49.22	-25.30	-30.30	13.70	10.50	7.80
8×8	-1.17	10.86	22.04	28.70	39.80	49.48	-8.60	-20.40	13.00	2.00	5.20
16×16	-1.58	10.66	21.16	30.49	39.92	48.77	-6.60	-11.60	-4.90	0.80	12.30
32×32	-1.82	10.84	21.85	30.61	39.94	48.49	-8.40	-18.50	-6.10	0.60	15.10
64×64	-1.80	11.10	21.33	30.75	39.87	48.56	-11.00	-13.30	-7.50	1.30	14.40

of the first-order predictor is comparatively less than that of the second-order predictor. Results of orders higher than the second are not shown, because performance deteriorates as the order increases.

Table 3: Performance based on different images for the first-order noncausal and causal second-order predictors in term of EL and ER

Type	Embedding Percentage											
	First-order noncausal linear predictor						Second-order noncausal linear predictor					
	0%	10%	20%	30%	40%	50%	0%	10%	20%	30%	40%	50%
EL	-1.46	10.64	20.75	30.02	40.28	49.62	-3.23	12.08	23.11	31.34	39.55	47.07
ER	–	-6.40	-7.50	-0.20	-2.80	3.80	–	-20.80	-15.55	-4.46	1.12	5.86

4.4 Performance for Different *k*

Table4 4 gives the results of dependency of the performance of extraction on the values of k for noncausal and causal predictors on average results. The result is taken for values of k from 1 to 9 with the increment 2 for block size of 16×16. Tables show that the performance of the proposed technique was better for higher values of k. It is because the predictor can give larger value of absolute error $|e(m,n)|$ when the predicted pixel is different from the neighboring pixels.

Table 4: Performance based on value of k for the first-order noncausal predictor in terms of EL and ER

Value of k	Embedding Percentage										
	EL						ER				
	0%	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
1	18.70	23.73	17.08	29.11	25.34	28.21	-137.30	29.20	8.90	146.60	217.90
3	1.92	11.27	23.32	32.45	31.11	52.08	-12.70	-33.20	-24.50	88.90	-20.80
5	1.48	12.83	21.57	36.55	37.23	48.81	-28.30	-15.70	-65.50	27.70	11.90
7	1.78	9.10	18.62	31.88	38.16	49.45	9.00	13.80	-18.80	18.40	5.50
9	-1.46	10.64	20.75	30.02	40.28	49.62	-6.40	-7.50	-0.20	-2.80	3.80

5 Conclusions

This paper presents a novel steganalytic technique of $\pm k$ steganography based on noncausal linear predictor using prediction coefficients obtained from a block of pixels in the image. The effectiveness of the proposed technique in estimating the length of the embedded message bits in the stego-image was verified using different types of images. It was found that noncausal predictor gives better results than causal predictor, and the first-order predictor gives better performance than the second-order. It is also found that the performance is better for higher values of k.

Bibliography

[1] Kawaguchi E.; Noda H.; Niimi M.; Eason R.O. (2003); A Model of Anonymous Covert Mailing System Using Steganographic Scheme, in *Information modelling and knowledge bases*, XIV, H. Yaakkola et al. (Eds), IOS Press, 81-85.

[2] Wang H.; Wang S. (2004); Cyber Warfare Steganography vs Steganalysis, *ACM Commun.*, 47: 76-82.

[3] Anderson R.J.; Pettitcolas, F.A.P. (1998); On The Limits Of Steganography, *IEEE Journal on Selected Areas in Communication*, 16(4): 474-481.

[4] Zhang X.; Wang S. (2006); Dynamically Running Coding In Digital Watermarking, *IEEE Signal Processing Letters*, 13(3): 165-168.

[5] Provos N.; Honeyman P. (2003); Hide And Seek: An Introduction To Steganography, *IEEE Security And Privacy*, 1(3): 32-44.

[6] Westfeld A.; Pfitzmann A. (1999); Attack On Steganographic Systems, *Proc. Information Hiding - 3rd Intel Workshop*, Springer Verlag, 61-76.

[7] Fridrich J.; Goljan M. (2000); Practical Steganalysis Of Digital Images - State Of The Art, Security And Watermarking Of Multimedia Contents IV, E.J. Delp III and P.W. Wong, editors, *Proc. of SPIE*, 4675, 1-13.

[8] Zhi L.; Fen S.A.; Xian Y.Y.A. (2003); LSB Steganography Detection Algorithm, *Proc. IEEE ISPIMRC*, 2780-2783.

[9] Zhang T.; Ping X. (2003) ; Reliable Detection Of LSB Steganography Based On The Difference Image Histogram, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3: 545-548.

[10] Mielikainen J. (2006); LSB Matching Revisted, *IEEE Signal Processing Letters*, 13(5): 285-287.

[11] Li X.; Yang B.; Cheng D.; Zeng T. (2009); A Generalization Of LSB Matching, *IEEE Signal Processing Letters*, 16(2): 69-72.

[12] Fridrich J.; Soukal D.; Goljan M. (2005); Maximum Likehood Destination Of Secret Message Length Embedded Using PMK Steganography In Spatial Domain, *Proc. Of IST/SPIE Electronic Imaging: Security, Steganography And Watermarking Of Multimedia Contents*, VII, 5681, 595-606.

[13] Balram N.; Moura M.F. (1996); Noncausal Predictive Image Codec, *IEEE Trans. On Image Processing*, 5(8): 1229-1242.

[14] Schroeder M.R.; Atal B.S. (1985); Code-excited Linear Prediction (CELP): High Quality Speech At Very Low Bit Rates, *IEEE Proc. ICASSP*, 937-940.

[15] Fang W.; Yagle A.E. (1994); Two-dimensional Linear Prediction And Spectral Estimation In Polar Raster, *IEEE Trans. On Signal Processing*, 42(3): 628-641.

[16] Jain A.K. (1989); *Fundamental of Digital Image Processing*, Prentice Hall, Information and System Science Series.

[17] Singh K. M. (2012); Vector Median Filter Based On Noncausal Linear Prediction For Detection Of Impulse Noise From Images, Inderscience *International Journal of Computational Science and Engineering*, 7(4): 345-355.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 9(5):633-643, October, 2014.

A Modified Ant Colony Algorithm for Traveling Salesman Problem

X. Wei, L. Han, L. Hong

Xianmin Wei*, Liqi Han, Lu Hong
School of computer engineering, Weifang University,
Weifang 261061, P.R. China
weixm@wfu.edu.cn, liqi.han@wfu.edu.cn, sdkfz234@163.com
*Corresponding author: weixm@wfu.edu.cn

Abstract: Ant colony algorithms such as Ant Colony Optimization (ACO) have been effectively applied to solve the Traveling Salesman problem (TSP). However, traditional ACO algorithm has some issues such as long iterative length and prone to local convergence. To this end, we propose we embed ACO into Cultural Algorithm (CA) framework by leveraging the dual inheritance mechanism. Best solutions are evolved in both population space and belief space, and the communication between them is achieved by accept and influence operations. Besides, we employ multiple population spaces for parallel execution. Experiments show that the performance of our proposed algorithm is greatly improved.
Keywords: Traveling Salesman Problem (TSP), Ant Colony Optimization (ACO), Cultural Algorithm (CA).

1 Introduction

Traveling Salesman Problem (TSP) is a typical NP hard problem [1]. Given a number of cities and the distances between them, TSP aims to calculate the shortest tour for the salesman to visit each city exactly once and return to the starting point. Solving TSP problem can be beneficial to applications such as deployment of network devices [2], transportation [3] and traffic control [4].

Ant Colony Optimization (ACO) algorithm [5] is one possible solution for solving TSP problem. ACO simulates the foraging process of ants, where the routes with more pheromones are more likely to be selected. The process of applying ACO algorithm for TSP can be described as follows. Suppose there are ants and cities. The movement of each ant is to choose the next unvisited city to visit by certain rules, and meanwhile update the amount of pheromones on that route.

Although the advantages of ACO, such as positive feedback, distributed, parallel and self-organization, facilitate the solving of TSP problem, there still remain some issues. For example, at earlier iterations, the amount of pheromones is relative exile, so the accumulation of enough pheromones costs a long time. Besides, at later iterations, the positive feedback mechanism greatly increases the possibility of local convergence.

To deal with the above long iteration time and local convergence issues, we propose an improved algorithm in this paper. The general idea is to accelerate the evolution and revise the best solution at each iteration. To this end, we leverage the Cultural Algorithm (CA) [6] framework to embed the traditional ACO algorithm. Indeed, as indicated in [7, 8], it is feasible and efficient to adapt CA for optimize ACO. Our experiments prove that the dual inheritance mechanism of CA helps to improve the performance of ACO.

The remainder of this paper is organized as follows. Section 2 reviews related work. The proposed algorithm is discussed in Section 3. Empirical experiments of evaluating the improved algorithm are conducted in Section 4. Finally, the paper is concluded in Section 5.

2 Related work

Many efforts have been made to solve the problems of traditional ACO algorithm. For example, Dorigo et al. [9] proposed an ant system with positive feedback, distributed computation, and a constructive greedy heuristic. Bullnheimer et al. [10] proposed a new rank based version of ant system. Gutjahr et al. [11] proposed a graph based ant system. Gambardella et al. [12, 13] introduced Q-learning to ant system for solving TSP problem. Ciornei et al. [14] developed a hybrid algorithm with ant colony and genetic algorithm, so that the proposed algorithm can achieve faster convergence and better search capability for global best solution. Stutzle et al. [15] designed a Max-Min Ant System (MMAS) for better exploitation ability and accelerated the accumulating the pheromone of the best solutions. Walter and Thomas et al. proved the convergence of ACO algorithm [16–18].

Indeed, ACO algorithms have been applied to many problems such as the TSP problem. Zhou et al. [19] provided the theoretical convergence analysis on the traditional ACO, and proved the feasibility of applying ACO algorithm for TSP. Zhang et al. [20] proposed an adaptive heterogeneous multiple ACO to solve TSP, which used an evolution coefficient to evaluate the solutions obtained by the ant colony and then update the phenomenon. Tuba et al. [21] modified ACO by improving the pheromone correction mechanism, where the pheromone values for highly undesirable links are significantly lowered by a posteriori heuristic. Negulescu et al. [22] presented an adapted ACO that incorporates methods and ideas from genetic algorithms (GA) by simulating artificial ants with different behaviors as synthetic genes. Besides, ACO has been employed in many applications. Gajpal et al. [23] applied ACO to the route selection of tucks, and the search all the customers during the routing path using a multi-route search principle. Hu et al. [24] employed a new method to update the phenomenon to deal with the continuous optimization problems. Ghoseiri et al. [25] developed an algorithm to solve the dual-object shortest path problem with a multi-objective ACO algorithm. Gajpal et al. [26] proposed a modified ACO method for a vehicle routing problem. Secui et al. [27] applied ACO for optimal allocation of capacitor banks in electric power distribution networks.

Cultural algorithm (CA) was first proposed by Reynolds [6]. Then, Chung et al. [28] developed a new framework to embed the evolutionary programming into the population space to simulate the evolution process. Liu et al. [29] combined CA with particle swarm optimization algorithm and applied it to the numerical optimization problem. Ma et al. [30] applied CA to solve the TSP problem and utilized the idea from simulated annealing to ensure the accuracy and efficiency.

Unlike existing work, in this paper we design a modified ACO algorithm by leveraging CA framework, that is, to employ a dual inheritance mechanism into the traditional ACO algorithm. Besides, we apply the modified algorithm on TSP problem. Note that, for the consideration of parallelism, we use multiple population spaces.

3 Proposed algorithm

The disadvantages of the traditional ACO algorithm include slow convergence and prone to stagnation. Although there are some modifications over the traditional ACO version, most of them are built upon the traditional computing structure to determine the internal status of the algorithm and estimate the search capabilities of the ant colony through qualitative analysis. For example, the improvements are made by updating pheromones, changing evaporation coefficient or the amount of released pheromones in order to avoid stagnation.

Different from existing efforts, we embed the ACO algorithm into another framework, Cultural Algorithm (CA) [6], which could essentially improve the performance. The intuitive is that both algorithms are population based and share information through interactions among the

population. We perceive that the ACO algorithm can benefit from the dual inheritance of CA. Besides, the inherent nature of parallelism of ACO indicates that parallel algorithm can improve the efficiency due to the mature of hardware and software. Therefore, in this section, we propose a parallel ACO algorithm based on CA algorithm, and then apply it for solving the traditional TSP problem.

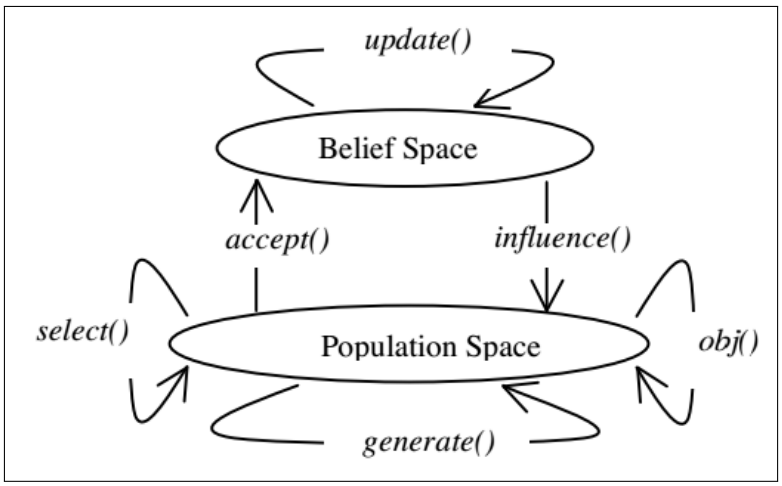


Figure 1: The framework of culture algorithm

CA simulates the evolution process of human society, which regards culture as the information carrier. While the culture is accepted by the whole population of the society, it can be used to direct the behavior of each individual in turn. There are two evolution spaces in CA algorithm: belief space, which is composed of the experience knowledge acquired during the evolution process; and population space, which is composed of the individuals. These two spaces communicate through specific protocols. The framework of CA is illustrated as Figure 2. Generally, the lower space contributes to the upper space periodically, and the upper space continuously evolves, which in turn influences the lower space. This is called dual inheritance mechanism.

As shown in Figure 2, the major operations of CA algorithm are `accept()` and `influence()`, while other operations are performed inside belief or population space independently. Therefore, it is feasible to embed other algorithms into CA framework by adding specific logics into belief and population spaces and implementing `accept()` and `influence()` operations between them. The parallel version of CA is illustrated as Figure 3, where exist multiple population spaces. Each population space performs parallel evolution, and then produces local best individuals for next generation. Then `accept()` operation is performed by all population spaces periodically to update belief space in a synchronous way. On the other hand, belief space conducts evolution as well, and calls `influence()` operation to direct the evolution process of each population space.

3.1 Designing population space

Let m be the number of ants, n be the number of cities $C = \{C_1, C_2, \ldots, C_n\}$, $d_{ij}(i, j = 1, 2, \ldots, n, \quad i \neq j)$ be the distance between any two cities, and τ_{ij} be the amount of pheromones from i to j. At the beginning, τ_{ij} is initialized as a constant, that is, $\tau_{ij}(0) = c$, where c is a constant. At t-th iteration, the pheromones on path i, j is notated as $\tau_{ij}(t)$. According to MMAS [12] algorithm, $\tau_{ij}(t)$ is restricted between $\tau_{\min}$ and $\tau_{\max}$ in order to avoid premature local convergence.

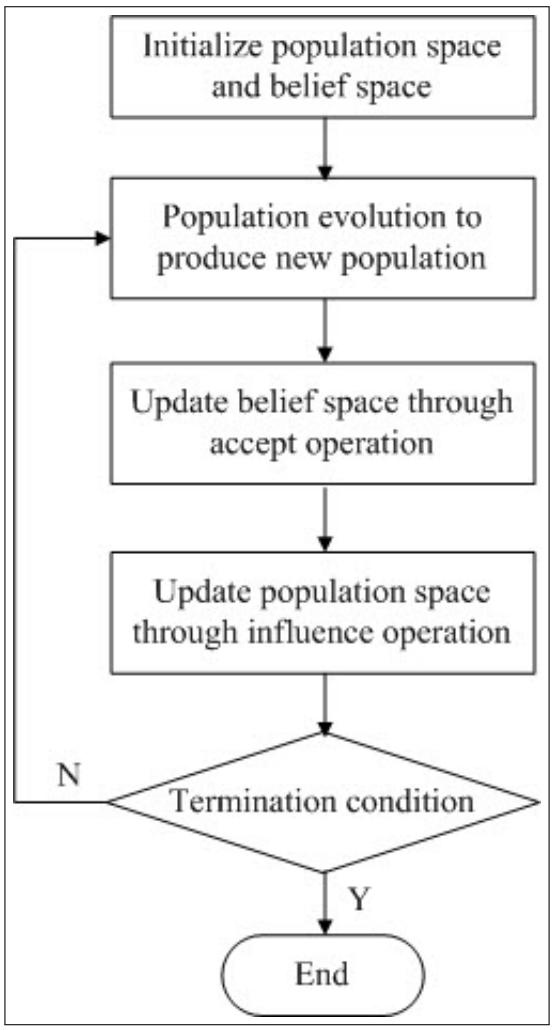


Figure 2: The process flow of culture algorithm

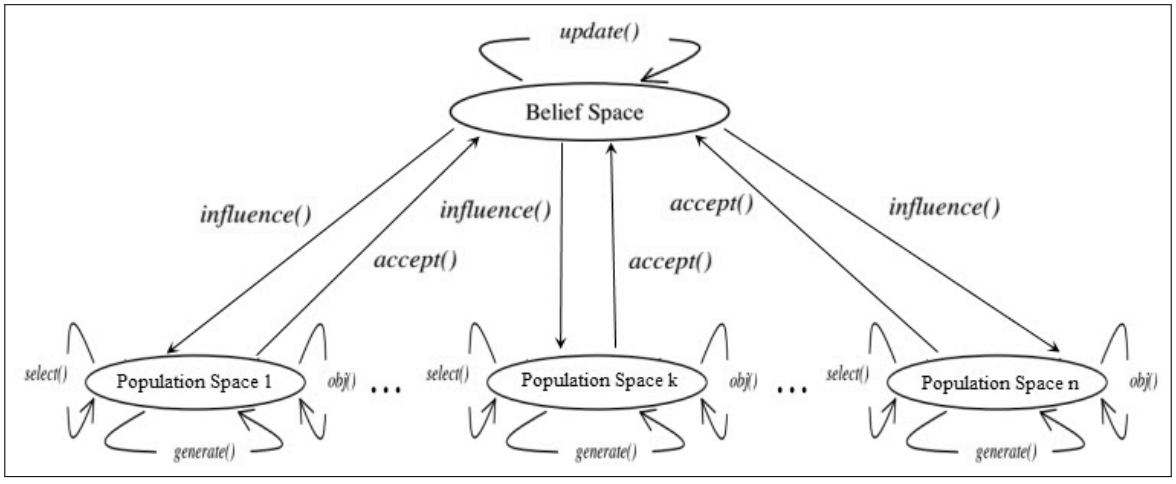


Figure 3: The framework of culture algorithm

$$\tau_{\max}(t) = \frac{1}{2(1-\rho)L(t)} + \frac{\sigma}{L(t)} \tag{1}$$

$$\tau_{\min}(t) = \frac{\tau_{\max}(t)}{20} \tag{2}$$

where $L(t)$ is the length of the best solution at iteration t, ρ is the evaporation coefficient of pheromones, and σ is the number of best solutions at iteration t.

Suppose $\eta_{ij}(t)$ denotes the heuristic information on path (i, j). The general idea is that ants would select closest path with larger amount of pheromones. Therefore, the probability of ant k transferring from city i to j is calculated as:

$$p_{ij}^k(t) = \begin{cases} \dfrac{\left[\tau_{ij}(t)\right]^\alpha \cdot \left[\eta_{ij}(t)\right]^\beta}{\sum\limits_{j \in \text{allowed}_k} \left[\tau_{ij}(t)\right]^\alpha \cdot \left[\eta_{ij}(t)\right]^\beta}, & j \in \text{allowed}_k, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where allowed_k denotes the available nodes to choose at the next step for ant k, α is the heuristic factor, meaning the importance of path with remaining pheromones, β is the heuristic factor for $\eta_{ij}(t)$, denoting the affect of heuristic information.

After ants iterate all the nodes, remaining pheromones are updated as follows:

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \sum_{k=1}^{m} \Delta\tau_{ij}^k(t) \tag{4}$$

where $1-\rho$ is the residual coefficient of pheromones, and $\rho \in [0, 1)$. $\Delta\tau_{ij}^k(t)$ is the amount of pheromones remaining on the path at current iteration for ant k, which can be calculated as:

$$\Delta\tau_{ij}^k(t) = \begin{cases} \frac{Q}{L_k}, & \text{if ant } k \text{ passes } (i, j) \text{ at current iteration}, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where Q is a constant, and L_k is the total length of ant k's tour.

3.2 Designing belief space

Belief space is responsible for updating knowledge, that is, to further optimize the best solutions provided by the population space, which are transferred through `accept()` operation. For TSP problem, we employ 3-OPT algorithm to evolve belief space. Suppose there exist any three nodes i, j, k, and current best solution is $C = \{C_s \cdots C_i C_{i+1} \cdots C_j C_{j+1} \cdots C_k C_{k+1} \cdots C_t\}$. As shown in Figure 4, if $d(C_i, C_{i+1}) + d(C_j, C_{j+1}) > d(C_i, C_j) + d(C_{i+1}, C_{j+1})$, then reverse sort is performed on $C = \{C_{i+1}, \ldots, C_j\}$; if $d(C_j, C_{j+1}) + d(C_k, C_{k+1}) > d(C_j, C_k) + d(C_{j+1}, C_{k+1})$, then reverse sort is performed on $C = \{C_{j+1}, \ldots, C_k\}$.

3.3 Accept operation

We accept a fixed ratio of all population spaces as the belief space, e.g. 20%. The `accept()` operation passes the local best solution of each population space R^* and the corresponding length of route L^* to the belief space.

When `accept()` operation is performed, the relative poor individuals are replaced by the current best one of each population space. Suppose t_{current} is the current iteration, t_{end} is the predefined maximum iterations, and t_{accept} is the iteration where `accept()` is called, which can be calculated as:

$$t_{\text{accept}} = \text{fix}\left(C_1 + \frac{t_{\text{current}}}{t_{\text{end}}} C_2\right) \tag{6}$$

where fix() is a rounding function, and C_1, C_2 are constants.

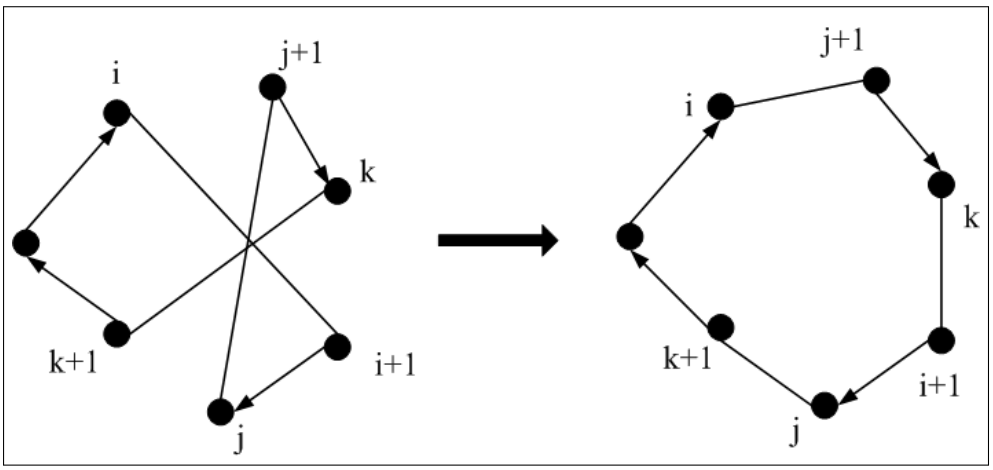


Figure 4: 3-OPT algorithm

3.4 Influence operation

After further optimization via 3-OPT in belief space, `influence()` operation passes the shortest path and its length to each population space periodically, and updates the global pheromones. Suppose the iteration where `influence()` operation is performed is $t_{\text{influence}}$, which can be represented as:

$$t_{\text{influence}} = \text{fix}\left(C_1 + \frac{t_{\text{end}} - t_{\text{current}}}{t_{\text{end}}} C_2\right) \tag{7}$$

where fix() is a rounding function, and C_1, C_2 are constants.

The improved algorithm has one belief space and multiple population spaces. Each space evolve independently, and the communication between them is achieved by `accept()` and `influence()` operations. The process flow of belief space is shown in Figure 5. In belief space, the update is achieved by 3-OPT algorithm for computing the shortest path. When `accept()` is called, the solution is updated by the local best from population space. When `influence()` is called, the global best solution in belief space is transferred to each population space. If the termination condition is satisfied, the algorithm is completed.

As shown in Figure 6, the update of population space is achieved by ACO algorithm. When `accept()` is called, the local best solution from each population space is updated up to the belief space. When `influence()` is called, the global best solution in belief space is transferred down to each population space. If the termination condition is satisfied, the algorithm is completed.

4 Experiment

The environment of our experiments is Intel 2.3 GHz, 2GB RAM, Windows 7 operating system, and MATLAB 7.0 for programming. The dataset is achieved from TSPLIB library . The parameters settings are as follows: $\rho = 0.5$, $\alpha = 1$, $\beta = 5$, $Q = 100$, $t_{\text{end}} = 200$. We use 4 population spaces, and the maximum number of evolution of both belief space and population space is 200. We compare our algorithm with the traditional ACO. Both algorithms are ran independently for 50 times, and the average execution time is supposed to the performance.

As shown in Table 1, we evaluate three examples from TSPLIB, i.e., eil51, berlin52 and st70. We can see that our proposed algorithm outperforms the typical ACO method for solving TSP problem. Moreover, the best solution provided by our method is much better than the known best.

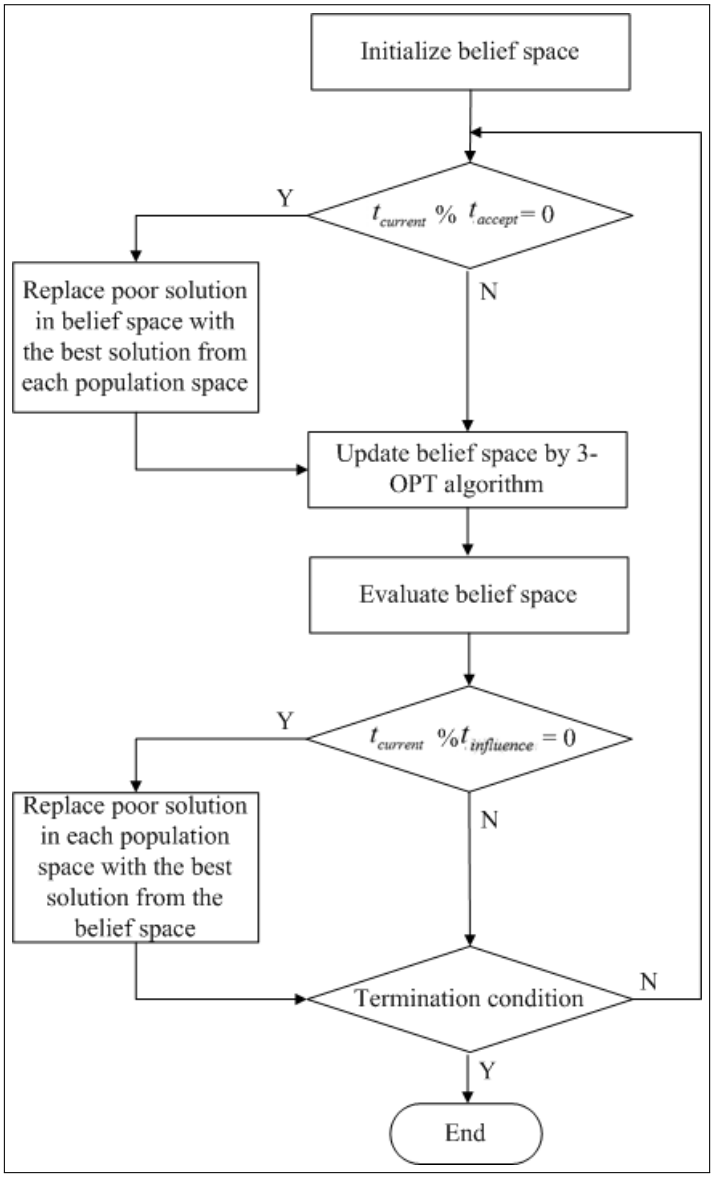


Figure 5: Flow chart of evolution of belief space

Table 1: Results of ACO and our algorithm with different datasets

Example	Example cities	Example solution	Example ACO	Our algorithm
eil51	51	426	428.87	412.12
berlin52	52	7542	7542	7088.34
st70	70	675	677.11	652.80

Besides, take eil51 as an example, we compare the best tour and convergence speed between our algorithm and traditional ACO for TSP in Figure 7. The blue line denotes the performance of ACO, while the green line is for our algorithm. The result shows that our algorithm can achieve shorter tour with rapid speed of convergence.

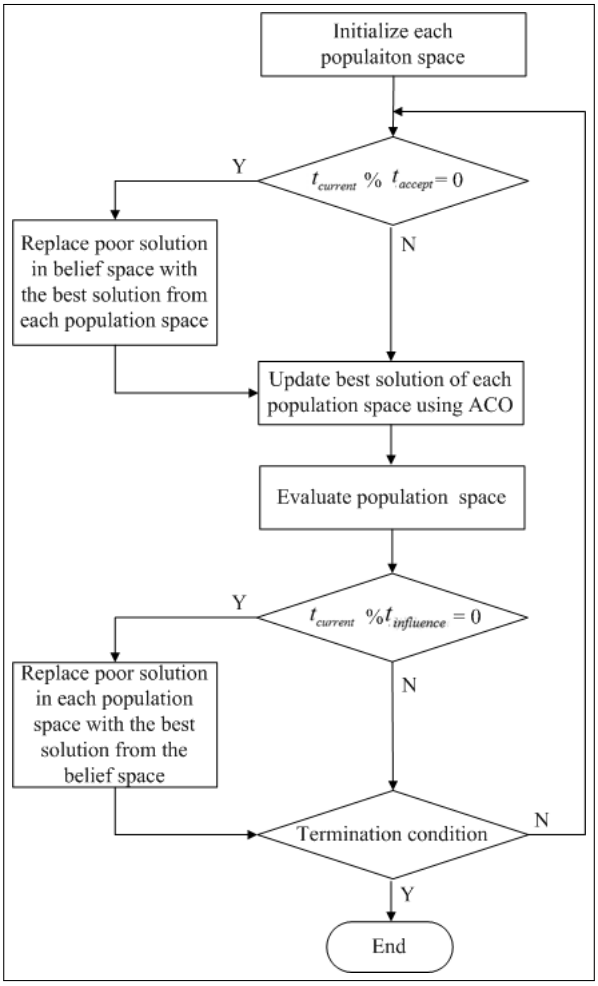


Figure 6: Flow chart of evolution of population space

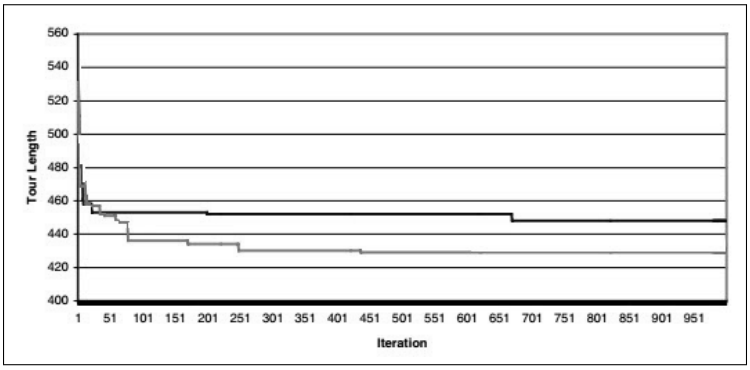


Figure 7: Evolution of best tour length of eil51 for ACO and proposed algorithm

5 Conclusions

In this paper, we propose a new ant colony algorithm and apply it to the typical traveling salesman problem. Specifically, we embed the traditional ACO algorithm into cultural algorithm framework, where the set of ants makes up the population space, and the best solutions found at specific iteration are updated up to the belief space. The evolution of population space is indeed based on ACO, and that of the belief space is based on 3-OPT algorithm for calculating the shortest path. The communication between population space and belief space is performed by accept and influence operations. Moreover, to minimum the execution cost, we design multiple population spaces for parallel execution. The experiments evaluate that our algorithm outperform the traditional ACO algorithm.

Acknowledgments

This work was partially supported by Shandong Natural Science Foundation (ZR2011FL006), Shandong Science and Technology Development Plan (2011YD01044), Shandong Spark Program (2012XH06005), and Weifang municipal Science and Technology Development Plan (201301050).

Bibliography

[1] Lenstra J., Kan A., Shmoys D. (1985), *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*, Vol. 3, New York: Wiley, USA.

[2] Xing G., Wang T., Jia W., Li M. (2008), Rendezvous Design Algorithms for Wireless Sensor Networks with a Mobile Base Station, *Proc. of the 9th ACM International Symposium on Mobile ad hoc Networking and Computing 2008*, ACM, 231–240.

[3] Feillet D., Pierre D., Michel G. (2005), Traveling Salesman Problems with Profits, *Transportation Science*, ISSN 0041-1655, 39(2): 188–205.

[4] Tariq M., Ammar M., Zegura E. (2006), Message Ferry Route Design for Sparse ad hoc Networks with Mobile Nodes, *Proc. of the 7th ACM International Symposium on Mobile ad hoc Networking and Computing*, Florence, Italy, 37–18.

[5] Dorigo M. (2006); Ant Colony Optimization and Swarm Intelligence, *Proc. of 5th International Workshop*, ANTS 2006, Brussels, Belgium, Vol. 4150, Springer.

[6] Reynolds R. (1994), An Introduction to Cultural Algorithms, *Proc. of the Third Annual Conference on Evolutionary Programming*, Singapore, 131–139.

[7] Gu J., Fan P., Song Q. (2010), Improved Culture Ant Colony Optimization Method for Solving TSP Problem. *Computer Engineering and Applications*, ISSN 1002-8331, 46(26):49–52.

[8] Liu S., Wang X., You X. (2009), A Cultural Ant Colony System for Solving TSP Problem. *Journal of East China University of Science and Technology (Natural Science Edition)*, ISSN 1671-4512, 35(2):288–292.

[9] Dorigo M., Vittorio M., Alberto C. (1996), Ant System: Optimization by A Colony of Cooperating Agents, *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 26(1) : 29–41.

[10] B. Bullnheimer B., Hartl R., Strauss C. (1997), A New Rank Based Version of the Ant System, *A Computational Study.*

[11] Gutjahr W. (2000), A Graph-based Ant System and Its Convergence, *Future Generation Computer Systems*, ISSN 0167-739X , 16(8): 873–888.

[12] Gambardella L., Dorigo M. (1995), Ant-Q: A Reinforcement Learning Approach to the Traveling Salesman Problem, *ICML*, 252–260.

[13] Dorigo M. , Luca M. (1996), A Study of Some Properties of Ant-Q, *Parallel Problem Solving from Nature-PPSN IV*, Springer, Berlin, Heidelberg, 656–665.

[14] Ciornei I., Elias K. (2012), Hybrid Ant Colony-genetic Algorithm (GAAPI) for Global Continuous Optimization, *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(1): 234–245.

[15] Stutzle T., Holger H.(1997), MAX-MIN Ant System and Local Search for the Traveling Salesman Problem, *Proc. of the 1997 IEEE International Conference on Evolutionary Computation, ICEC'97*, Indianapolis, IN, USA, 309–314.

[16] Gutjahr W. (2003), A Generalized Convergence Result for the Graph-based Ant System Metaheuristic, *Probability in the Engineering and Informational Sciences*, ISSN 0269-9648, 17(4): 545–569.

[17] Gutjahr W. (2003), A Converging ACO Algorithm for Stochastic Combinatorial Optimization, Stochastic Algorithms: Foundations and Applications. Springer Berlin, Heidelberg, 10–25.

[18] T. Stützle, M. Dorigo (2002), A Short Convergence Proof for A Class of Ant Colony Optimization Algorithms, *IEEE Trans. Evolutionary Computation*, ISSN 1089- 778X, 6(4): 358–365.

[19] Y. Zhou (2009), Runtime Analysis of an Ant Colony Optimization Algorithm for TSP Instances, *IEEE Trans. Evolutionary Computation*, ISSN 1089-778X, 13(5): 1083–1092.

[20] Zhang P., Jie L., Ling X. (2010), An Adaptive Heterogeneous Multiple Ant Colonies Algorithm, *Journal of Intelligent Systems*, ISSN 2191-026X, 19(4): 301–314.

[21] Tuba M., Jovanovic R. (2013); Improved ACO Algorithm with Pheromone Correction Strategy for the Traveling Salesman Problem. *International Journal of Computers Communications & Control*, 8(3):477–485.

[22] Negulescu S.C., Dzitac I., Lascu A.E. (2010); Synthetic genes for artificial ants. Diversity in ant colony optimization algorithms.*International Journal of Computers Communications & Control*: 5(2):216-223.

[23] Gajpal Y., Prakash A. (2009), An Ant Colony System (ACS) for Vehicle Routing Problem with Simultaneous Delivery and Pickup, *Computers & Operations Research*, ISSN 0305-0548 , 36(12): 3215–3223.

[24] X. Hu, Z. Jun, L. Yun (2008), Orthogonal Methods Based Ant Colony Search for Solving Continuous Optimization Problems, *Journal of Computer Science and Technology*, ISSN: 1000-9000, 23(1): 2–18.

[25] Ghoseiri K., Behnam N. (2010), An Ant Colony Optimization Algorithm for the Bi-objective Shortest Path Problem, *Applied Soft Computing*, ISSN: 1568-4946, 10(4): 1237-1246.

[26] Gajpal Y., Prakash L.A. (2009), Multi-ant Colony System (MACS) for a Vehicle Routing Problem with Backhauls, *European Journal of Operational Research*, ISSN: 0377-2217, 196(1): 102–117.

[27] Secui D.C., Dzitac S., Bendea G.V., Dzitac I. (2009); An ACO Algorithm for Optimal Capacitor Banks Placement in Power Distribution Networks, *Studies in Informatics and Control*, 18(4): 305–314.

[28] Chung C.J., Robert G.R. (1998), CAEP: An Evolution-based Tool for Real-valued Function Optimization Using Cultural Algorithms, *International Journal on Artificial Intelligence Tools*, ISSN: 0218-2130, 7(3): 239–291.

[29] Liu S., Wang X., You X.M. (2007), Cultured Differential Particle Swarm Optimization for Numerical Optimization Problems, *ICNC 2007. Proc. of Third International Conference on Natural Computation*, Haikou, Hainan, China, 642–646.

[30] Ma J. (2008), Research on Cultural Algorithm for Solving Routing Problem of Mobile Agent, *The Journal of China Universities of Posts and Telecommunications*, ISSN:1002-1310 , 15(4): 121–125.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 9(5):644-654, October, 2014.

A Knowledge-based Telemonitoring Platform for Application in Remote Healthcare

W. Zhang, K. Thurow, R. Stoll

Weiping Zhang*, **Kerstin Thurow**
University of Rostock
Celisca, Center for Life Science Automation, Rostock 18119, Germany
*Corresponding author: weiping.zhang@uni-rostock.de
kerstn.thurow@uni-rostock.de

Regina Stoll
University of Rostock
Institute for Preventive Medicine, University Medical Center, Rostock 18055, Germany
regina.stoll@uni-rostock.de

Abstract: Telemonitoring systems have been shown to greatly reduce medical costs while improving the quality of medical care. Today, the main factors restricting the development and popularization of Telemonitoring systems include scalability and compatibility. The challenge for the remote healthcare lies in the variety of heterogeneous medical sensors which need to be dynamically removed or added to the environment according to the health care needs. This paper presents the design for an ontology-based context model and related middleware that provides a reusable and extensible application platform for Remote Healthcare. We designed the ontology context model to describe physiological parameters, medical tasks and the patients personal profile. Developers may extend the ontology model by considering new requirements as needed.
Keywords: telemonitoring, ontology, knowledge-based System, Context-Aware application

1 Introduction

Increasing life expectancy, due to medical progress, and decreasing birth rates lead to a change in the population structure of Germany. As illustrated in Fig. 1, the percentage of people over the age of 65 is steadily increasing while the number of people under the age of 25 is declining [1]. This results in continuously rising costs for healthcare, especially for elderly people. One approach to this problem is the utilization of telemonitoring systems to enable a relocation of the therapy to the patients home, thus shortening the expensive periods of hospitalization.

A Medical Telemonitoring System can be defined as a technological means for sending remote physiological information and medical signals through a communication network to a monitoring center for analysis and diagnostics. The system generally include three components: a monitoring center, medical sensors, and a communication network connecting the two components.

However Telemonitoring systems are still in the early stages of development and currently face many challenges. Compared to a general monitoring environment, a remote healthcare environment has the following features: (1) Dynamistic: the remote healthcare environment usually requires a number of medical sensing devices and, unlike air conditioning which is fixed in an environment, some medical sensors need to be removed or added to the environment according to the health care needs; (2) Complexity: in the healthcare environment, a wide variety of medical terminology are used, and the relationships between them are very complex [2].

In order to solve these problems, we designed a Mobile Context and Ontology-based Reasoning/Feedback (MCOM) health-monitoring system which monitors a patients health status using a Smartphone. On the Smartphone, a Context-Aware Middleware is developed to connect

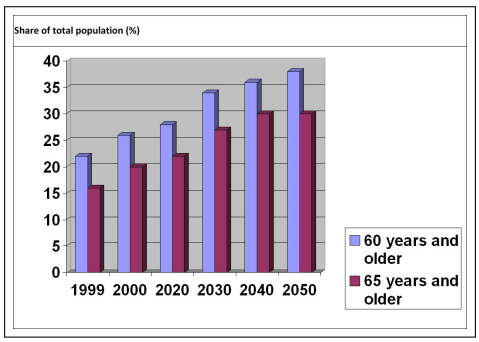


Figure 1: Projected Percentage of Elderly People in German Population

different types of medical sensors. In addition, MCOM uses an ontology-based context model to process and determine a patients health status.

2 Healthcare Context-Aware Middleware (HCAM)

2.1 Context-Aware Middleware

The first problem above mentioned can be solved by establishing a Context-Aware Middleware [3]. Context-Aware computing refers to a general class of mobile systems that can sense their physical environment, and adapt their behavior accordingly [4]. In order to extricate remote medical applications from tedious sensor data acquisition and management, this paper proposes a Healthcare Context-Aware Middleware (HCAM). With HCAM, the upper-layer applications would not be concerned about acquiring the context data, but only about the business logic itself.

As shown in Table.1, applications based on HCAM are divided into three layers. HCAM separates the collection of context information and the development of aware applications. It provides a Subscriber API (i.e. Subscriber Application Programming Interface [5]) for the developer of the upper-layer application; HCAM is then responsible for completing the data collection of various kinds of sensors.

Table.1 Healthcare Context-Aware Middleware	
Context aware applications	
Healthcare Context-Aware Middleware	Subscriber API
	Pub/Sub framework
	Provider API
	Data collection
All kinds of sensors	

2.2 Ontology Modeling

The second problem can be solved using the Ontology Context Model and Reasoning. Ontology can be defined as, an explicit and formal specification of a shared conceptualization [6] and provides a kind of shared vocabulary, including object type or concepts, their properties, and relationships existing in special fields. In this work, it is used to define information and relations, including sensor data, medical tasks, control strategies, and a patients personal domain knowledge. A medical task, like an Emergency Alert, can be triggered by Context Information reasoning based on the analysis of physiological parameters.

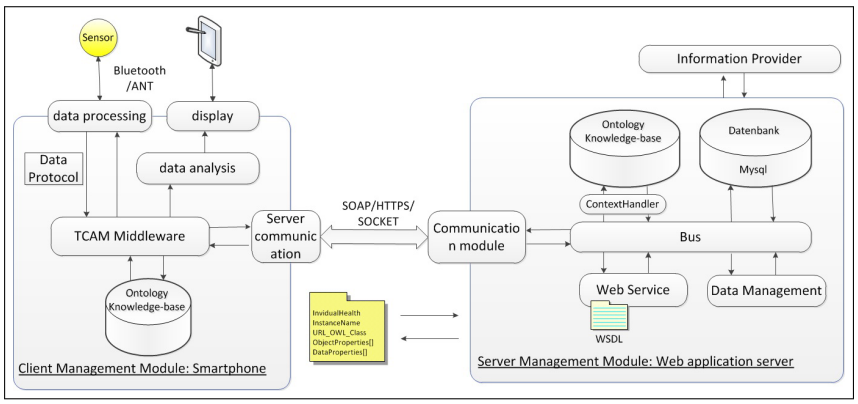


Figure 2: Telmonitoring System Architecture: Management Modules and Communication link

2.3 System structure

This work aims at describing the ontology-based context model and middleware that have been developed in order to ease the development of Context-Aware applications for remote healthcare [7]. The system architecture is shown in Fig.2. From a physical perspective, it is based on a C/S architecture which includes the sensor terminal, Smartphone client terminal, and remote server terminal. From a functional perspective, the whole system can be divided in to three modules: Client Management Modules (CMD), Server Management Modules (SMD), and communication mechanisms.

1. The core module at the Smartphone client is the HCAM Middleware. It is responsible for communicating with various sensors and performing data collection. Other core components include data processing, analysis, and display modules plus server communication modules. It should be noted that, unlike other Telemonitoring systems, the communication between client and server is bidirectional. In addition to sending data to a server, the client terminal can also synchronize information with the server. The design concept of the client terminal is to offload as much processing as possible to the server, so that the Smartphone client operates only as gateway to display and transfer data.

2. On the Server-side, the core component is the knowledge-base, based on ontology. Other components on Server include a communication module that is responsible for transmitting information between the server and the client terminal, a data management module, plus a web service module. These modules communicate through a task bus.

3. As shown in Fig.2, the communication between CMD and SMD is implemented through HTTP/HTTPS, SOAP [18] and SOCKET. HTTP is used in synchronous XML workflow description; SOAP is used to transmit ontology instance information; SOCKET connection is used in high real-time demand, for example, transmitting ECG (1024Byte/Sec) and Acceleration (200Byte/Sec) data.

2.4 HCAM Structure

The HCAM proposed in this paper is designed and developed based on the Android system for Mobile Phone. Its architecture is shown in Fig.3. According to the previously mentioned three-layer structure, HCAM include these three layers, i.e. context data acquisition, context data transmission and distribution, etc. HCAM first uses the context provider to collect the original context data provided by the sensors, then transmits the data to the upper layer applications via

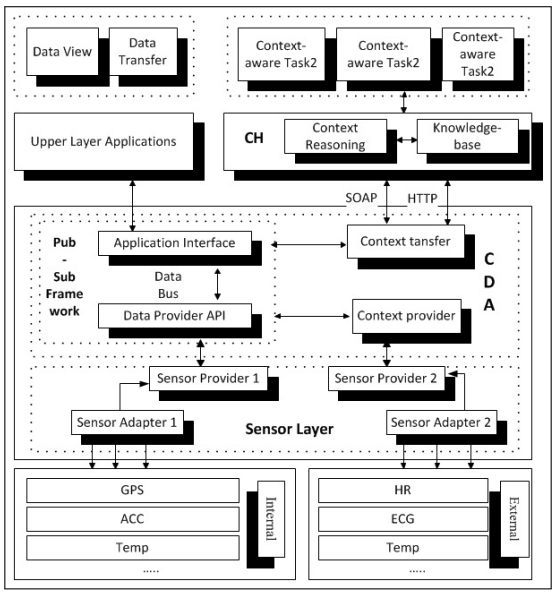


Figure 3: HCAM Architecture

the Pub/Sub framework. The Pub/Sub framework [8] completes the context data transmission and distribution. The following are the core components of the HCAM .

1. **Sensor Layer** All the original context of physiological data are derived from the sensor, which is the basis of any Context-Aware application. This paper divided sensors into two categories; internal sensors (located in the Smartphone) and external medical sensors.

2. **Data Provider** Data provider (Internal) 1 and data provider (External) 2 are shown in Fig. 3. A data provider usually exists in the form of a plug-in and is written using Provider API. Once the providers are completed, the Provider API will transmit the data to the upper sensing applications. During the actual realization process, data provider 1 and data collector 1 are usually combined and exist as many different system plug-ins. However, data provider 2 and the data package resolver, are combined and exist as an independent plug-in. Under these circumstances, the data provider can directly transmit the data, read by the data collector, and the data analyzed, by the data package resolver, to the upper sensing applications via Data-Bus [9].

3. **Context Discovery and Access (CDA)** CDA consists of the context provider, Pub/-Sub framework, and context transfer. The context provider serves as a context discovery protocol to connect different sensors. The Pub/Sub framework collects context information from the context provider, and the context transfer uploads the data to the central server

4. **Context Handler (CH)** CH reasoning is accomplished through the context instance from the client sensor data. The context information uses an inference engine based on conditions to conduct reasoning via OWL/RDF [10] modeling (Knowledge-base). Once the conditions are met, the context handler module will send information to the Context-Aware task and monitoring server.

3 Knowledge-base design

Before processing the Context Information, we first need to construct the Context Model. In an ubiquitous medical environment, a large number of medical devices and domain terms are used. With the advancement of technology, new domain terms can be joined very quickly. In addition, in the rescue and medical care process, the pathological conditions of a patient usually cannot be identified directly; medical professionals must filter and integrate information from a variety of different sources. Therefore, the context model permits not only adding the description of a new concept, but also supports the context reasoning.

By using ontologies and various reasoning methods, the system can process the sensor data gathered in a more parameters to create a method by which the engine can infer the patients health status [11].

The context knowledge-base has two parts. The first part stores the ontology and their instances in the system. The other part provides a reasoning interface and context discovery, i.e., how to inquire and to add modules. Ontology instances may be pre-defined in the profile or may be retrieved from the device and other Webservice. The pre-defined profile is dynamically imported into the system at the start of system operation. The context information also includes some fixed information from a specific time; for example, the patients blood type, emergency telephone number, clinical records, etc. Fig.4 shows a UML diagram of the context ontology specialized for the Patient Personal Domain. In the Patient Personal Domain, relevant context items include Measurements, Patient Profile, Device, and Task. These data can be used by the system to automatically deduce the patient health status and detect possible medical tasks.

The Class Patient, (see Fig. 4), constitutes the core of the ontology. Each defined patient will be configured as an instance of this Class. One patient personal profile instance is generated and associated for one particular patient, but this patient could have more than one patient profile instance associated to him [12]. A specialization of the Measurements Class is added for each specific physiological parameter which is gathered from a sensor (e.g., HeartRate Class). Measurement values are used to determine patient health status by comparing these values with a set of parameter ranges (MeasurementsRange Class) or results of the specialized Data-processing-algorithm (like the Stress-assessment algorithm in [13]). When a measured value falls outside the limits, a task (e.g., Alarm Task) of the corresponding level is then triggered. For Example, Class HeartRate (see Fig. 4) is a subclass of the Measurements Class ,and in this case, there are three alert rangeslow, medium, and high which can be modified by the Expert to adapt to patients' specific needs. In the case of HeartRate, the three ranges indicate HeartRate is too low, Normal or too high. And the model can be easily extended to include further levels. The hasCurrentValue attribute contains the most recently received sensor value from the mobile client.

This system uses OWL to construct the model and process it using Sample Semantic Web Rule Language [14]. The reason for using the context to construct the model is that, it can not only express the context information, but can also use the inference engine based on conditions to conduct inference to the subtle context to get higher level information [2] The context models have the following forms: (Subject, Predicate, Value)

- Subject belongs to S*: set of subject names. a patient, a place , a Smartphone, or an Emergency Task for example.
- Predicate belongs to P*: the set of Predicate names, the patients current body temperature for example.
- Value belongs to V*: the overall value of the Subject.

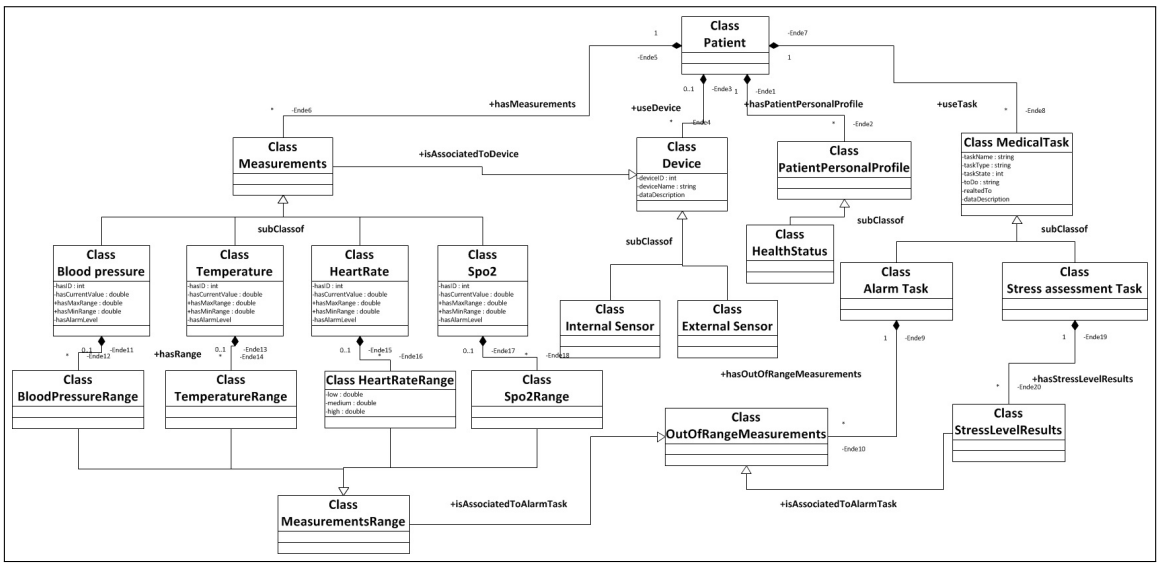


Figure 4: Patient Remote Healthcare Domain Ontology UML Representation

For example, we have a patient with ID 001, body temperature 37°C and use a PDA to connect to the monitoring center, then information would be recorded as follows:

- (PDA, online, true), (Patient001, Temperature, 37)

Table 2 shows a few example rules, which the system uses to predict the sensor-data-specific task and Table 3 shows basic example policy we defined for each Emergency Task level [7]. The HeartRateHigh rule assigns the Emergency Task level when the heart rate is abnormally high. The hasMinRange and hasMaxRange properties specify a normal heart rate in each instance of the HeartRate Class. If a patients heart rate is greater than the specified normal maximum, then the system sets the Heartrate alarm and starts the Emergency Task .

Table 2: rules for the systems alarm management	
RULE	DESCRIPTION
HeartRateHigh	(?patient rdf:type HeartRate) , (?par1 hasCurrentValue ?v1) , (?par1 hasMaxRange ?Max) greaterThan(?v1, ?Max) , − > (?taskstate hasAlarmLevel HES)
HeartRateLow	(?patient rdf:type HeartRate) , (?par1 hasCurrentValue ?v1) , (?par1 hasMinRange ?Min) , greaterThan(?v1, ?Min) , − > (?taskstate hasAlarmLevel LES)

Table 3: Alarm Task Notification Policy Examples	
Alarm TASK LEVEL	NOTIFICATION POLICIES
Low Alarm Task (LES)	sensor alarm sms to patient relative
Medium Alarm Task (MES)	sensor alarm sms , mail and call to relative
High AlarmTask (HES)	sensor alarm message to emergency operator call to relative

Unlike common ontology systems, in this paper the medical task is also included in the ontology; the task will invoke specific activities. There are two advantages in doing this: the first advantage is that the specific details of the task and the sub-actions are divided. Under these circumstances, the description of the task doesnt need to be changed when the device information, used by a sub-activity, changes. The other advantage is that the description of task by ontology, can make inference to the changes that should be made by the task; the inference processing depends on the task status, context information, and the custom rule. At the root node, is the design of a type of named task and a specific task level taking task. After inference is completed and the instances of task type exist, SOAP objects are sent to the task response components and then to the client terminal, including all task information. We use the following attributes to describe a task (see experiment):

1. TaskName: the name of the task;
2. TaskType: The type of task and function description information;
3. Todo: describes the current sub-activities of the task;
4. TaskState: indicates the current state of a task. There are three values to be chosen: -CLOSE. , -INIT and -OPEN and they respectively indicate the task has not been started, or the task needs to be initialized or is started for use.
5. RelatedTo: indicates which specific function in the ontology the task is related to, and what kind of device instances the task is related to.
6. DataDescription: data description describes the data type and format that needs to be collected by the task. The information is stored in the task profile and uploaded when the system is operational.

3.1 Simulation

In this section we will use an emergency assistance task as a practical example to describe the above described design pattern: how new devices are added to the ontology knowledge-base and medical treatment begins via conditional reasoning.

The specific scenario for an emergency assistance task is as follows: An anomaly has been detected for a patient suffering from heart disease or high blood pressure; based on system reasoning, emergency assistance is requested. According to this task profile, the response time of the task is one minute, meaning that medical personnel must respond within one minute, react to the patients condition and initiate corresponding measures. After the response, the tasks mode is set to CLOSED. If the patient's condition remains abnormal, the task is started again.

The first step is to add new devices: In the experiment we used the wearable multi-parameter sensor Equivital01 produced by (Hidalgo Ltd., Cambridge).

When a new device is connected to the HCAM sensor adapter, context information must be converted into ontology instances before being sent to the knowledge base. The HCAM context handler inquires whether classes with identical names already exist in the knowledge-base. If not, a new class is added with the new $< device >$. The following example shows the class of an Equivital device being added:

- $< owl : Classrdf : ID = Equital >$
  $< rdfs : subClassOfrdf : resource = *Device/ >$
  $< owl : disjointWithrdf : resource = */ >$
  $< /owl : Class >$

Then a new class is added which acts as subclass of Measurement. Below the classes HeartRate and BloodPressure are added for the Equital device.

- $< owl : Classrdf : about = *HeartRate >$
  $< rdfs : sublassOfrdf : resource = *Measurement/ >$
  $< owl : disjointWithrdf : resource = */ >$
  $< /owl : Class >$
  $< owl : Classrdf : about = *BloodPressure >$
  $< rdfs : sublassOfrdf : resource = *Measurement/ >$
  $< owl : disjointWithrdf : resource = */ >$
  $< /owl : Class >$

With the establishment of the above-mentioned ontology, it is ensured that the device description file and the ontology name match. When the HCAM context provider receives data from the device, it looks for the parameters in its own device files and device task files and creates an instance. The above-mentioned Equivital device might, for example, receive event messages from the HCAM detailing HeartRate-hasCurrentValue = 60(Beats/Minute), BloodPressure-hasCurrentValue = 120(systolic, mmHg). In this case the practical example would look as follows:

- $< HeartRaterdf : about = *EquitalHeartRate >$
  $< hasCurrentValuerdf : datatype = double > 60.0 < /hasCurrentValue >$
  $< /HeartRate >$
  $< BloodPressurerdf : about = *EquitalBloodPressure >$
  $< hasCurrentValuerdf : datatype = double > 120.0 < /hasCurrentValue >$
  $< BloodPressure >$
  $< Equitalrdf : about = *Equital001 >$
  $< measurementsrdf : resourcee = *EquitalHeartRate/ >$
  $< measurementsrdf : resource = *EquitalBloodPressure/ >$
  $< /Equital >$

After the context information is converted, the context handler will send the aforementioned information to the knowledge-base.

The second step is to execute the task.
When the middleware is executed, the system automatically generates an instance for every task class, which is saved in the knowledge-base. The important information in these task instances is as follows:

- $< MedicalTaskrdf : about = *EmergencyTask >$
  $< taskType > External < /taskType >$
  $< taskState > CLOSE < /taskState >$
  $< todordf : resource = *AlarmTask/ >$
  $< todordf : resource = *PrepareEmergencyOperator/ >$
  $< /MedicalTask >$

With the previous scenario in mind, we can define the specific conditions for the emergency task to be triggered: e.g. the Heart Rate of the patient drops below 40, or the blood pressure (systolic) exceeds 170. Reasoning may be carried out according to the rules of the language itself, for example:

- $(?a?p?b), (?prdf : subPropertyOf?q) - > (?a?q?b)$

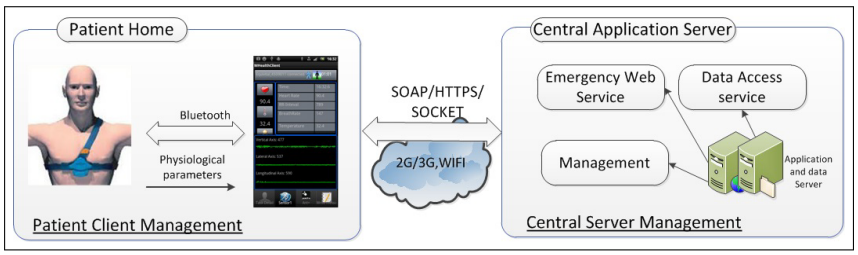


Figure 5: mHealth Prototype System Architecture

User defined rules may also be used in the reasoning process, e.g.:

- $(?patient, EquitalBloodPressure001, vl), GE(?vl, 170), (?patient, EquitalBloodPressure, vl), LE(?vl, 40) -> (?patient, healthStatus, danger)$
- $(?patient, healthStatus, danger), (?Patient, healthStatus, movementfail), (EmergencyTask, taskState, CLOSE) -> (EmergencyTask, taskState, OPEN)$

3.2 Deployment and Execution

The deployment configuration includes at least two Context Manager nodes as shown in Fig.5: We used a SonyEricsson-Xperia smartphone as Client Manager (CM) node, connected with the medical sensor module Equivital [15] via Bluetooth; and a Server Manager (SM) node, which is deployed on a Central Application Server, in the domain of the organization which is responsible for the integration and delivery of Emergency Web Task.

The Home Smartphone hosts the Client and communicates with the Central Application Server via 2G/3G/WiFi. The Central Application Server hosts the Server SM and other assistance and information task applications. These tasks include: a Emergency Task as Webservice, Data management, and a web application which offers web-based access to health operators via mobile device.

4 Results and Discussion

Fig.6 shows the sensor data for a patient. The graphic curve represents any changes in the data for a given day. Date and time can be changed easily, access a patients historical data; this data are stored in a central server database. The alarm is normally triggered automatically by context reasoning and, in some cases, the medical professionals can manually send a message to the client. Directly above the line chart in Fig.6, a message window is located. Medical professionals can send a message to the patient's Smartphone through it.

In this technical approach, the system structure improves the scalability and interactivity of telemonitoring systems; In our current work, we improve the reliability of the system by adding layers of redundancy, in case a particular system or function fails. In addition, we created a back-up system on the mobile client to save sensor data locally, in the event of a loss (or lack) of internet connection. Security issues can be divided into two parts: one is the security of local data transferred from the sensor to the mobile client; another is the security of remote data transferred from the mobile client to the central server. On the client side we implement the new 128-bit lightweight cryptographic algorithm Hummingbird-2 [16] cipher to encrypt the sensor data with java code. On the server side, we use Secure Socket Layer (SSL) to encrypt vital data and utilize 3P-AAA-SPs [17], Authorization and Accounting Task Providers to authenticate the client access.

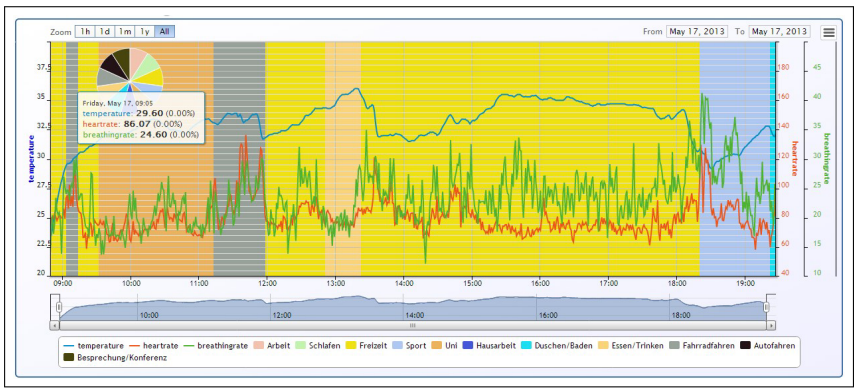


Figure 6: mHealth Prototype System Architecture

This paper designs an innovative Telemonitoring platform based on context and ontology technology. A Context-Aware Middleware is developed to support the scalability of the telemedical system; new sensor device and data formats can be integrated into the system conveniently. In the context reasoning module, ontologies are used to construct target medical tasks and circumstances. This paper, however, does not address the handling of context conflicts in the process of constructing ontologies and reasoning conditions. Further research will focus on the reliability and energy consumption problem of remote medical systems. Concerning the energy consumption problem, some new ultra low power transmission protocols, like BLE and ANT, have appeared that can greatly improve the sustainability of remote monitoring. Other challenging tasks will include the development of an intelligent error discovery and system recovery mechanism to maintain the stability of the system over time.

Bibliography

[1] Sitzungsberichte der Leibniz-Sozietat, Band 90 (2007); *Theoria cum praxi. Funf Jahre Leibniz-Institut fur interdisziplinare Studien e.V. (LIFIS)*, Trafo Verlag, Berlin, 217-228.

[2] Cheng J., Rao R. (2009); A Context-Aware middleware for pervasive healthcare. *Computer Applications and Software*, 50-53.

[3] Baldauf M., Dustdar S., Rosenberg F. (2007); A survey on Context-Aware systems. *International Journal of AdHoc and Ubiquitous Computing*, 2(4) : 263-277.

[4] Kuna M., Kolaric H., Bojic I., Kusek M., Jezic G. (2011); Android/OSGi-based Machine-to-Machine Context-Aware system. *Proc. of the 11th Int. Conf. on Telecommunications, ConTEL*, 95-102.

[5] Papapanagiotou I., Falkner M., Devetsikiotis M. (2012); Optimal functionality placement for multiplay task provider architectures. *IEEE Transactions on Network and Task Management*, 9(3): 359-372.

[6] Studer R., Benjamins V.R, Fensel D. (1998); Knowledge Engineering:principles and methods. *Data and Knowledge Engineering*, 25:161-197.

[7] Paganelli F., Giuli D. (2007); An ontology-based context model for home health monitoring and alerting in chronic patient care networks. *Proc. 21st Int. Conf. on Advanced Information Networking and Applications Workshops/Symposia*, 838-845.

[8] Wang N., Yang Z., Yang Y.(2011); Based on event-driven and task-oriented architecture business activity monitoring design and implementation. *Int. Conf. on System Science, Engineering Design and Manufacturing Informatization, ICSEM*, 241-245.

[9] Xiong L., Niu J., Zhang J., Chen F., Shen P.(2011); Context Sensing Middleware Based on Mobile Devices. *Journal of Chinese Computer Systems*, 1170-1174.

[10] Armas R., Cuenca G., Horrocks I.(2012); MORe: Modular combination of OWL reasoners for ontology classification. *Lecture Notes in Computer Science*, 1-16.

[11] Benlamri R., Docksteader L. (2010); MORF: A mobile health-monitoring platform, *IT Professional*, 12(3):18-25.

[12] Lasierra N, Alesanco A, Garcia J. (2010); An ontology approach to manage individual patient profiles in home-based telemonitoring scenarios. *Proc. of the IEEE/EMBS Region 8 Int. Conf. on IT Applications in Biomedicine, ITAB*, 1-4.

[13] Kumar M., Weippert M., Vilbrandt R., Kreuzfeld S., Stoll R. (2007); Fuzzy evaluation of heart rate signals for mental stress assessment. *IEEE Trans Fuzzy Syst*, 15:791-808.

[14] Horrocks I. et al. (2004); SWRL: A Semantic Web Rule Language Combining OWL and RuleML, World Wide Web Consortium (W3C), www.w3.org/Submission/SWRL.

[15] Neubert S., Arndt D., Thurow K., Stoll R. (2012); Mobile real-time data acquisition system for application in preventive medicine, *Telemedicine and e-Health*, 16:504-509.

[16] Engels D., Saarinen M.J.O., Schweitzer P., Smith E.M. (2012); The hummingbird-2 lightweight authenticated encryption algorithm. *Lecture Notes in Computer Science*, 19-31.

[17] Le X.H, Khalid M., Sankar R., Lee S. (2011); An Efficient Mutual Authentication and Access Control Scheme for Wireless Sensor Networks in Healthcare. *Journal of Networks*, 355-364.

[18] Tekli J.M, Damiani E., Chbeir R., Gianini G. (2012); SOAP processing performance and enhancement. *IEEE Transactions on Services Computing*, 3: 387-403.

Author index