Borworn Papasratorn
Kittichai Lavangnananda
Wichian Chutimaskul
Vajirasak Vanijja (Eds.)

Communications in Computer and Information Science 114

Advances in Information Technology

4th International Conference, IAIT 2010
Bangkok, Thailand, November 2010
Proceedings

Springer

Communications
in Computer and Information Science 114

Borworn Papasratorn
Kittichai Lavangnananda
Wichian Chutimaskul
Vajirasak Vanijja (Eds.)

Advances in Information Technology

4th International Conference, IAIT 2010
Bangkok, Thailand, November 4-5, 2010
Proceedings

Springer

Volume Editors

Borworn Papasratorn
Kittichai Lavangnananda
Wichian Chutimaskul
Vajirasak Vanijja
King Mongkut's University of Technology Thonburi
School of Information Technology
126 Pracha-U-Thit Rd., Bangmod, Thungkru, Bangkok 10140, Thailand
E-mail: {borworn;kitt;wichian;vachee}@sit.kmutt.ac.th

Library of Congress Control Number: 2010937473

CR Subject Classification (1998): D.2, I.2.6, I.2, H.4-5, C.2, E.3

ISSN 1865-0929
ISBN-10 3-642-16698-9 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-16698-3 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180 5 4 3 2 1 0

Preface

It was our intention to organize IAIT 2010 as the place where researchers and industrial practitioners could share their work and achievements in information technology, both theoretical and in application. This is the fourth time we have organized IAIT to serve as a venue to foster collaboration among industry and academia from different parts of the world. A lot has been achieved since the launch of the first IAIT in 2005; however, many challenges remain to be addressed in the years to come. IAIT 2010 drew attendance from leading professionals in both industry and academes, as in IAIT 2009. In addition to the relevant program offered by IAIT 2010, the conference provided an excellent environment to meet peers in the IT profession, build relationships, and exchange lessons learned. During the conference, participants presented and discussed the latest topics in information technology, ranging from technical knowledge and experimentation to future strategic directions.

As the Internet continues to reach out even further, almost everything in our daily life is accessible via IP addresses, our societies need to recognize that we need new knowledge to make this advancement become a benefit to all life in this world. It is almost impossible to gain this new knowledge without contributions from our colleagues, working in various aspects of information technology. Some times, new knowledge found in one area can help simplify difficult task in another. Together we can make our world better by creating applications from information technology.

Many thanks go to everyone who submitted presentations for this event. The additional time that the presenters have taken to document and share their ideas with us is greatly appreciated. Our thanks also go to many people for helping make this conference happen, including our Advisory Committee, keynote speakers and editor from Springer.

We are pleased that we were able to host the conference at a time when we commemorated the 15th anniversary of the School of Information Technology (SIT), KMUTT. It was the good time to share this special occasion.

We thank those who participated in this year's IAIT and making it possible by their attendance.

September 2010 Borworn Papasratorn

Organization

IAIT 2010 was organized by the School of Information Technology, King Mongkut's University of Technology Thonburi.

Executive Committee

Honorary Chair	Borworn Papasratorn (Thailand)
Program Chair	Wichian Chutimaskul (Thailand) Kittichai Lavangnananda (Thailand)
Advisory Committee	Prasert Kanthamanon (Thailand) Roland Traunmuller (Austria) Joaquin Vila-Ruiz (USA) Billy Lim (USA) Masatoshi Yoshikawa (Japan)

Program Committee

Organization Committee	Vajirasak Vanijja (Thailand) Kriengkrai Porkaew (Thailand) Vithida Chongsuphajaisiddhi (Thailand) Ekapong Jungcharoensukying (Thailand) Suree Funilkul (Thailand)

Referees

R. Alcock
C. Arpnikanondt
J.H. Chan
Y. Chen
V. Chongsuphajaisiddhi
N. Churcher
W.Chutimaskul
S. Funilkul
O. Halabi
S. Kalayavinai
A. Kamiya
K. Lavangnananda
W. Lertlum
K. Lertwachara
H. Marzi
F. Masaru
P. Mason
P. Mongkolnam
C. Nukoolkit
K. Porkaew
O. Rojanapornpun
A. Srivihok
G. Stylianou
U. Supasitthimathee
M. Suzuki
Y. Suzuki
V. Vanijja
S. Wangpipatwong
T. Wangpipatwong
N. Waraporn
B. Watanapa
F. Yang
C. Yu

Table of Contents

Evaluating a Novel Cellular Automata-Based Distributed Power Management Approach for Mobile Wireless Sensor Networks

Sepideh Adabi[1], Sahar Adabi[1], and Ali Rezaee[2]

[1] Islamic Azad University, North Tehran Branch, Tehran, Iran
[2] Islamic Azad University, Science and Research Branch, Tehran, Iran

Abstract. According to the traditional definition of Wireless Sensor Networks (WSNs), static sensors have limited the feasibility of WSNs in some kind of approaches, so the mobility was introduced in WSN. Mobile nodes in a WSN come equipped with battery and from the point of deployment, this battery reserve becomes a valuable resource since it cannot be replenished. Hence, maximizing the network lifetime by minimizing the energy is an important challenge in Mobile WSN. Energy conservation can be accomplished by different approaches. In this paper, we presented an energy conservation solution based on Cellular Automata. The main objective of this solution is based on dynamically adjusting the transmission range and switching between operational states of the sensor nodes.

Keywords: Cellular Automata, Energy Conservation, Mobile Wireless Sensor Network.

1 Introduction

A wireless sensor network is a special kind of network where the nodes can sense, actuate, compute and communicate with each other using multi-hop communication. The mobile wireless sensor network that is considered will constitute a multi-hop network.

Our network is homogeneous, that is all the nodes have a common structure furthermore a dynamic approach is considered in which after deployment, a certain node can change its location and move with same speed in different directions.

It is almost very difficult to change or recharge batteries in sensor nodes, since sensor nodes are commonly small and distributed in a large area [2]. Therefore, there have been research efforts on energy- conservating techniques in sensor networks.

An important issue in maximizing the lifetime of a network is minimizing the global energy which is consumed through sensing, communicating and movement. To minimize energy consumption, a common technique is to put a number of sensors in the sleep mode and put the others in the active mode for sensing and communicating tasks. According to three fundamental properties of Cellular Automata: *Parallelism, Locality* and *Homogeneity* [3], a certain energy-conservating technique based on CA that we will call it ECCA (Energy-Conservating based on CA) has been proposed in

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 1–10, 2010.

static large wireless sensor networks[1], but the idea behind the technique uses in mobile wireless sensor networks too. The rule of the mentioned CA is derived from a probing-based density control algorithm [4]. ECCA has just focused on achieving a longer network lifetime through turning off those nodes which are performing a redundant monitoring task in specific periods of time, while for increasing the life time of a network dynamic range adjustments can also be considered. That is, if we set the transmission ranges in such a way that global energy consumption is reduced, we expect the network's lifetime to be increased.

The paper is structured as follows. Section 2 gives important notifications and preliminaries. Section 3 discusses the key concepts of Cellular Automata (CA) Section 4 describes the mathematical energy consumption model and section 5 introduces the three-phased algorithm. Section 6 compares our solution with ECCA through simulation. Finally, our concluding remarks and possible future directions are presented in section 7.

2 Preliminaries and Important Attributes

The grid where the nodes are deployed is implemented as a bi-dimensional matrix. For a matrix of dimension *L*, two classes which are achieved from experimental results are defined for the number of deployed nodes (this is just because to consider different initial activated sensor nodes which are deployed over the region):

Class1	*Class2*
$\frac{L}{2}$ <Number of deployed nodes<= $\frac{(L^2-L)}{2}$	$\frac{(L^2-L)}{2}$ <Number of deployed nodes<= L^2

Each existing sensor node represents a cell of our cellular automata. Note that a cell will be assigned to a node if and only if the node occupies more than 55% of the cell's space. Each sensor is associated with the following attributes:

1) *State*. This attribute can have two values, 1 represent *active* mode and 0 represent *stand-by* mode.

Active nodes do processing, transmitting/receiving data packets and monitoring neighborhood region tasks. While a certain region is being monitored by two or more active nodes, a single node can switched to stand-by mode in order to save energy. Although it will listen to processing signals to be aware of the certain status of its neighbors.

2) R_{ca}. The *neighborhood radius* of the Cellular Automata.

3) *(i,j)*. The geographical position of each node that is set in the beginning of the simulation and change during network life time.

4) *Number of neighbors (sens-neigh)*. This attribute is determined by the number of existing nodes in the transmission range *(*R_{using} *)*.

5) *Sensing range* (R_{sense}).In fact the sensed area is the area within *sensing range* of each node.

6) *Node speed*. Mobile nodes can move in any possible directions with fixed speed which is equal to 1m/s.

7) *Timer (T_i).* Which will be calculated as illustrated in Equation (1):

$$T_i = ((\alpha \times D) * Ts) \quad \forall i \in \{All\ \ alive\ \ sensor\ \ nodes\} \tag{1}$$

Where D is the *node density* which equals to n/L^2, n is the number of nodes that are located in the grid space, T_s is the *time slot* which could be set to desired value and α is a random value between D and D/2. We use timer to prevent the unnecessary executions of the three-phased algorithm in each time slot to save energy.

3 Definition of Cellular Automata

As described in [5], Cellular Automata (CA) is decentralized, discrete space-time systems that can be used to model physical systems. Cellular Automata are formally defined as quadruples (d, S, N, f). The integer d is the *dimension* of the working space, obviously it is possible to create one, two or more dimensions automata. The dimension which is considered for this paper is two. S= {0, 1,…, s-1} is called the set of *states.* The neighborhood N= (n_1,…,n_v) is a ν-tuple of distinct vectors of Z^d. The n_i's are the relative positions of the neighbor cells with respect to the cell, the new state of which is being computed. The states of these neighbors are used to compute the new state of the center cell. The *local function* of the cellular automata f: $S^v \leftrightarrow$S gives the local transition rule. A *configuration* is a function from Z^d to S. The set of all configurations is C=S^{Z^d}. The *global function* A of the cellular automaton is defined via f as follows:

$$\forall c \in C, \forall i \in Z^d, \mathrm{A}(c)(i) = f(c(i+n_1),...,c(i+n_v))$$

4 Mathematical Model of Energy Consumption

4.1 Energy Consumption for Receiving and Transmitting Data

The energy declination function for each sensor node is calculated through a linear function. For a simplified power model of radio communication [6,7], the consumed energy in a transmission is:

$$E_t = (e_t + e_d r^n)\beta \tag{2}$$

Where e_t is the energy/bit which is consumed by the transmitter electronics, and e_d accounts for energy dissipated in the transmit op-amp. Both e_t and e_d are properties of the transceiver used by the nodes, r is the used transmission range. The parameter n is the power index for the channel path loss of the antenna. β is the bit rate of the radio and is a fixed parameter in our paper.

Also a fix amount of power is required to capture the incoming radio signal which is calculated as Equation (3):

$$E_r = (e_r \beta) \tag{3}$$

Where, e_r is the energy/bit which is consumed by the receiver electronics used by a node. Typical values for these attributes is described in [8].

Furthermore processing powers in Active state and Stand-by state are 0.0165J and 0.00006J respectively [9]. For evaluating the average amount of communicational energy the transmission power (based on Equation (2)) and reception power (based on Equation (3)) are combined to introduce a new average of consumed energy through Equation (4):

$$E_{ave} = \left(0.02 + R^2_{u\sin g} \times 10^{-4}\right) \tag{4}$$

Therefore the average of consumed energy in each mode can be indicated by Equation (5):

$$\begin{aligned} \textit{Active} \text{ state: } E_{active} &= \left(E_{ave} + 0.0165\right) \textbf{\textit{AND}} \\ \textit{Stand-by} \text{ state: } E_{s\tan d-by} &= \left(6\times 10^{-5}\right) \end{aligned} \tag{5}$$

4.2 Energy Consumption According to Move in Possible Directions

It is assumed that the nodes move randomly in all possible directions. According to the structure of most common sensors and for the sake of easy computation, each vertical or horizontal movement consumes 0.01J, while each diagonal movement consumes 0.014 J in each meter.

5 Proposed Algorithm

In this work *Borland Delphi Enterprise* program is used to simulate the three-phased algorithm. The pseudo code of the three-phased algorithm can be seen in *algorithm* 1, *algorithm* 2 and *algorithm* 3. In the simulator each cell of the grid can be occupied by a sensor node.

The three-phased algorithm focuses on achieving a longer network lifetime through performing dynamic range adjustments and turning off those nodes that are performing a redundant monitoring task in specific periods of time. First, T_s is set to desired value which is acceptable, then *algorithm1* is applied to calculate amount of energy consumption in each state.

Then, each node sets it's R_{using}*(transmission range for performing tasks)* to maximize transmission range of the applied sensor device. Following, each sensor node tunes the transmission range R_{using} *(algorithm2).*

Algorithm2 is based on a four- phase protocol which is executed at each sensor node periodically:

Phase 1. Finding-neighbor (R_{ca}, *i, j)*

This function calculates the number of existing nodes in neighborhood region based on R_{ca}. The geographical position of a node is determined by i and j.

Phase 2. Calculating R_{min}

Let R_{min} denote the minimum transmission range of each node. In the second phase, all nodes calculate R_{min} simultaneously and independently as it's illustrated in Equation (6):

$$R_{min} = \frac{(sensor.MaxEnergy - (D \times R_{max})) \times R_{init})}{E_{init}} \tag{6}$$

Algorithm1 **Node Scheduling**

```
While network is alive do
   For each cell do
     If node is alive (node's residual energy > threshold) then
        If state is active then
          1.1. Decrement node's energy (Active value + average
               routing packet value)
          1.2. Decrement energy for movement
          1.3. R-calculation
          1.4. Verification
         Else
          1.1. Decrement node's energy (Stand-by value)
          1.2. Decrement energy for movement
          1.3. R-calculation
          1.4. Verification
         end if.
      end if.
   end for.
end while.
```

Where *sensor.MaxEnergy* is the current maximumenergy, E_{init} is the maximum battery capacity and R_{init} is the transmission range according to E_{init}. Furthermore R_{max} is the maximum transmission range which corresponds to E_{max} which is determined by the physical layer and radio characteristic.

Algorithm2 **R-Calculation**

```
Set R_cs=R_max.
Finding-neighbor (R_cs , i, j).
D= sens-neigh / L^2
Calculate R_min.
Calculate R_ave.
Set R_cs =R_ave.
Finding-neighbor (R_cs , i, j).
D= sens-neigh / L^2
Calculate R_using.
Set each node timer (T_1).
```

Phase 3. Calculating R_{ave}

In the third phase, all nodes calculate R_{ave} (the average of $R_{min}$and R_{max}) simultaneously and independently as it's illustrated in Equation (7):

$$R_{ave}=\frac{1}{2}(R_{min}+R_{max}) \tag{7}$$

Phase 4. Calculating R_{using}

Calculating transmission range R_{using} as it's illustrated in Equation (8):

$$R_{using} = \min\left(\sqrt{\frac{n_{ave}}{n_{max}}} \times R_{current-using}\ , R_{max}\right) \tag{8}$$

Assume that $R_{current-using}$ is the current transmission range, sens-neigh (1) is the number of existing nodes in transmission range $R_{current-using}$ and sens-neigh(2) is the number of existing nodes in transmission rangeR_{ave}, therefore n_{ave} and n_{max} are calculated as it's illustrated in Equation (9):

$$n_{max} = \text{sens-neigh}_{(1)} - 1 \quad \textbf{AND} \quad n_{ave} = \text{sens-neigh}_{(2)} - 1 \tag{9}$$

Each node will use R_{using}as the neighborhood radius of the Cellular Automata in *algorithm3*. By consulting CA's rule (for switching to new state), R_{using} guarantees the connectivity.

Indeed, *algorithm3* is executed to check the state of each sensor node and switch between states in case it is necessary.

Each node consults CA's rule to retrieve a new state. As mentioned before there are two possible states for each node: *active* (turn on) and *stand-by* (turn off). To choose which node should be turned on/off we must observe the fact that, if only there are two or more active neighbors in the transmission range (R_{using}) at the moment of verifying (*algorithm3*), the node must change its state to the stand-by mode to save energy or the state won't be changed [1]. Each node repeat the mentioned procedure until the energy of the node be less than predefined threshold, to ensure about its state. If there are less than two or more active neighbors at the moment of verifying (*algorithm3*), the node must change its state to the active mode or the state won't be changed.

In ECCA, neighborhood radius in CA is equal to 1, but in three-phased algorithm the neighborhood radius is determined by transmission range (R_{using}), therefore greater numbers of nodes are set to the stand-by mode. Through this method we achieved better energy conservation.

6 Performance Analysis

In order to characterize the behavior of our proposed algorithm the *Borland Delphi Enterprise* program is used. The measured performance of the proposed algorithm was compared with the idea which is proposed in ECCA [1] in mobile wireless sensor networks. Five metrics for wireless sensor networks have been used:

- **i.** *Battery remaining*: The sum of residual energy on all nodes in a specific period of time.
- **ii.** *Fragmentation:* The number of times in which an active node can not find any other active ones in its *using transmission range (R_{using})* in a specific period of time.
- **iii.** *Connectivity:* Which is based on the number of connections in a specific period of time.
- **iv.** *Coverage*: The percentage of the area that is monitored by the sensing range of all active nodes in a specific period of time.
- **v.** *Active nodes*: The number of nodes which are set to active state in a specific period of time.

The results have been gained through 50 times of executions by averaging.T_s, $E_{init}$and $R_{max}$are set to 200ms, 0.9J and 6m respectively. The sensing range (R_{sense}) of the applied sensor is equal to 1m. The scenario is defined as follow:

Our grid size equals to 260m×260m; the numbers of sensors vary according to two predefined classes. In Fig.1-5 the results are indicated in both classes of node deployment. The average results which are gained through 50 times of executions are used to draw the curves.

Algorithm3 **Verification**

```
if node's energy is less than threshold then
   node is dead
else
  if T1 is zero then
Consult CA's rule to retrieve new state
else
    Decrement timer
  end if.
end if.
```

a) Battery remaining

The total amount of network's energy which is given in Fig. 1 shows the effect of using three-phased algorithm on a network's lifetime. The utilization of three-phased algorithm increases network's lifetime up to four times, this is just because the numbers of nodes in stand-by mode are increased through dynamical transmission range adjustment and determine the neighborhood radius based on transmission range (R_{using}) instead of the fixed value of 1 which is applied in ECCA.

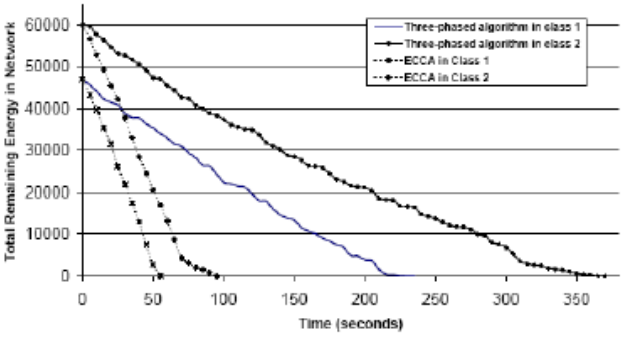


Fig. 1. Total amount of energy of the network in proposed algorithm vs. ECCA

b) Fragmentation

Fragmentation vs. time for ECCA and three-phased algorithm is illustrated in Fig. 2. This event (fragmentation) may occur when nodes are going to turn-off, but fragmentation count of nodes are higher than static wireless sensor networks due to node movement. Although a great number of nodes will be switched to stand-by mode in three-phased algorithm, less fragmentation is guaranteed according to Equation (8). Therefore, three-phased algorithm produces up to 20% less fragmentation than ECCA.

c) Connectivity

In this paper we defined a function to calculate the degree of connectivity of a network which is based on the number of connections in a specific period of time. That is, connectivity is defined as fragmentation's complement. The experimental results confirm this definition, in ECCA the degree of network connectivity is approximately 20%, but the three-phased algorithm achieves 43% of connectivity, as one can be observed in Fig.3.

d) Coverage

Observing Fig. 4, one can compare the coverage which is achieved by three-phased algorithm with the coverage which is gained through ECCA. Obviously week achieve a longer time of convergence through dynamical transmission range adjustment, furthermore the average of coverage is approximately over 80% during network's lifetime.

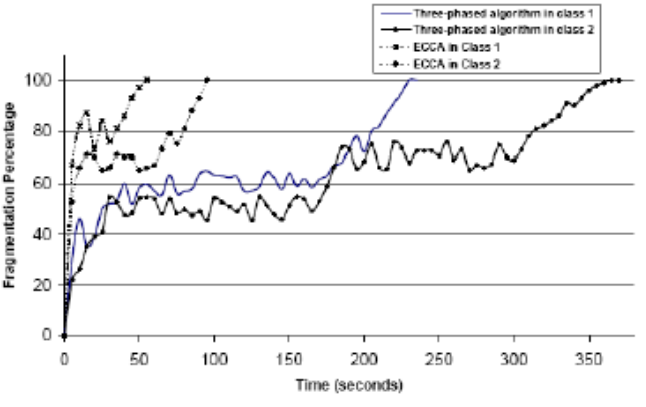


Fig. 2. Fragmentation in proposed algorithm vs. ECCA

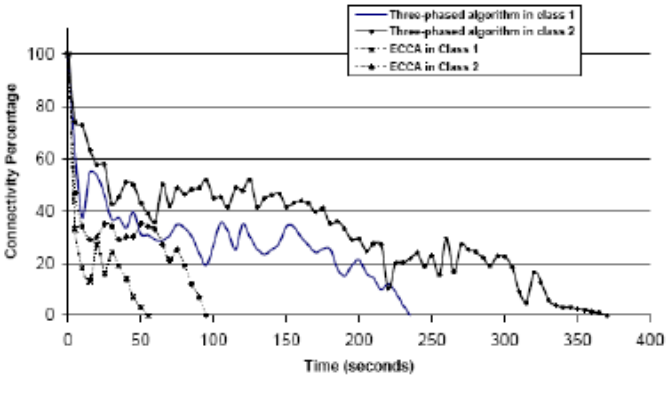


Fig. 3. The network connectivity in proposed algorithm vs. ECCA

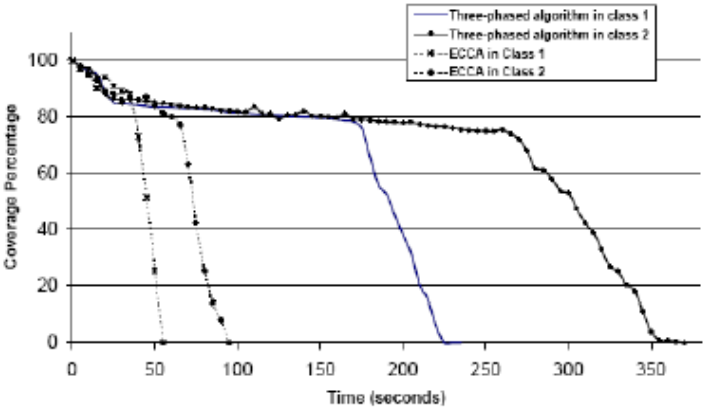


Fig. 4. The network coverage in proposed algorithm vs. ECCA

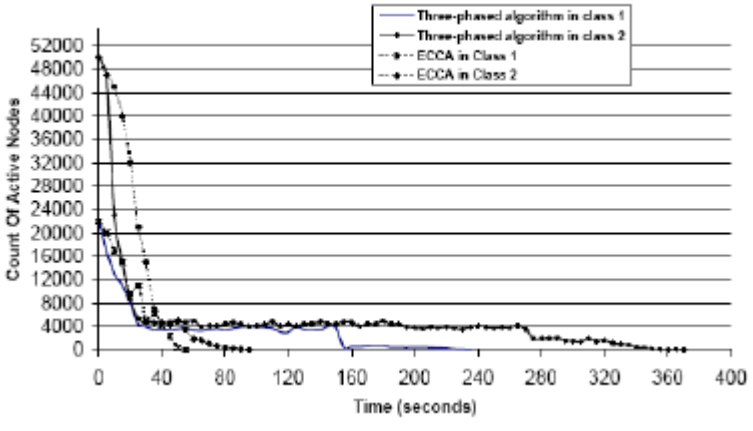


Fig. 5. The network active nodes in proposed algorithm vs. ECCA

e) Active nodes

The relation between the number of active nodes and time in ECCA and three-phased algorithm is given in Fig. 5. One can observe that, in order to save energy, greater numbers of nodes are set to stand-by mode instead of active mode, and therefore a longer network's lifetime is achieved.

7 Conclusion

In this paper, we proposed an energy-conservating three-phased algorithm in Mobile Wireless Sensor Networks which is based on Cellular Automata (CA). The three-phased algorithm focuses on achieving a longer network lifetime through turning off

those nodes which are performing a redundant monitoring task in a specific period of time and performing dynamic range adjustments. Our simulation proves that, through imposing suitable transmission range a great amount of energy will be saved in comparison with ECCA.

Future works can focus on extending the proposed algorithm in heterogeneous sensor network, which is all the nodes do not have a common structure.

References

1. Cunha, R.O., Silva, A.P., Loreiro, A.A.F., Ruiz, L.B.: Simulating Large Wireless Sensor Networks Using Cellular Automata. In: IEEE Proceedings of the 38th Annual Simulation Symposium (ANSS 2005), pp. 323–330 (April 2005)
2. Ye, W., Heidemann, J., Estrin, D.: Medium Access Control With Coordinated Adaptive Sleeping for Wireless Sensor Networks. IEEE/ACM Transaction on networking 12(3), 493–506 (2004)
3. Rucker, R., Walker, J.: Introduction to CelLab, May 10 (2002), `http://www.fourmilab.ch/cellab/`
4. Ye, F., Zhong, G., Lu, S., Zhang, L.: Energy Efficient Robust Sensing Coverage in Large Sensor Networks, Technical Report (2002)
5. Durand, B., Formenti, E., Grange, A., Róka, Z.: Number conserving cellular automata: new results on decidability and dynamics. Discrete Mathematics and Theoretical Computer Science AB, 129–140 (2003)
6. Min, R., Bhardwaj, M., Ickes, N., Wang, A., Chandrakasan, A.: The hardware and the network: total-system strategies for power aware wireless microsensors. In: Proceedings of IEEE CAS Workshop on Wireless Communications and Networking Pasadena, USA, pp. 36–42 (September 2002)
7. Bhardwaj, M., Garnett, T., Handrakasan, A.P.: Upper bounds on the lifetime of sensor networks. In: Proceedings of ICC 2001, Helsinki, Finlsnd, vol. 3, pp. 785–790 (June 2001)
8. Heinzelman, R., Chandrakasan, A., Balakrishnan, H.: Energyefficient communication protocol for wireless microsensor networks. In: Proceedings of HICSS 2000, Hawaii, USA, vol. 2, pp. 4–7 (January 2000)
9. Ruiz, L.B., Braga, T.R.M., Silva, F., Nogueira, J.M.S., Loureiro, A.A.F.: Service Management in Wireless Sensor Network. In: Operation and Management Symposium (LANOMS 2003), Foz do Iguaçu, PR, Brazil, pp. 55–62 (September 2003) ISBN: 85-902104-2-1

Effects of Feature Selection Using Binary Particle Swarm Optimization on Wheat Variety Classification

Ahmet Babalık, Ömer Kaan Baykan, Hazim İşcan, İsmail Babaoğlu, and Oğuz Fındık

Selçuk University, Department of Computer Engineering, 42075, Konya, Turkey
{ababalik,obaykan,iscan,ibabaoglu,oguzf}@selcuk.edu.tr

Abstract. In this article, classification of wheat varieties is aimed with the help of multiclass support vector machines (M-SVM) and binary particle swarm optimization (BPSO) algorithm. For each wheat kernel, 9 geometric and 3 color features are obtained from the digital images which are belong to 5 wheat type. Wheat types are classified using M-SVM. In order to increase the reliability of the classification process, 2 fold cross validation approach is implemented and this process repeated 250 times. Average classification accuracy is obtained as 91.5%. With the aim of increasing the classification accuracy and decreasing the process time, descriptive features are decreased by BPSO algorithm and reduced from 12 to 7. Average of classification accuracy is obtained as 92.02% using 7 features obtained from reduction with BPSO.

Keywords: Binary particle swarm optimization, support vector machine, wheat classification.

1 Introduction

Grain and grain products are the basic source of human beings. Shape, texture and color which belong to grain kernels are very important criteria which are used in quality and type inspection. Besides, visual reviews which are used in quality control are takes long time and tiring, and also, results can change from person to person. To remove this negation, quality control systems based on image processing techniques are developed in last years. [1]

Grain type identification is important from the aspect of giving price and process. Machine vision systems are used in classification and definition of grain [1,2]. Visen et al. have classified five grain types with artificial neural networks (ANN) by using color and texture feature [3]. Zapatoczny et al. have used morphological features and statistical approaches in order to classify the barley types [4]. Choudhary et al. have classified grain types using statistical techniques by assessing morphological, color and texture features [5]. Tahir et al. have used color and texture feature to detect grain moisture [6].

Wheat is an important commercial product which has operation in grain industry. Researchers suggested different image processing based systems to detect wheat types and their quality. Ramalingdam et al. have assessed morphological features of bulk grain to detect wheat moisture rates by using statistical approach [7]. Neethirajan

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 11–17, 2010.

et al. have used X-ray and transparency imaging to identified wheat vitreousness [8]. Wang et al. studied to determine wheat vitreousness by using ANN. Utku and Köksel have classified 31 different bread wheat with the help of statistical approaches using geometrical features belong to kernels [10]. Shouche et al. have classified 15 different wheat types with the help of statistical approaches by using 15 geometrical features [11]. Dubey et al. have classified 3 wheat type with ANN by using 45 morphometric features obtained from bulk wheat image [12].

The number of descriptive features belongs to grain, which obtained from digital images, effects classification accuracy and training time. In order to reduce training time and increase performance of the classification different pre-processing techniques (such as principal component analysis and fuzzy logic) have used [2, 13, 14]. In last years, SVM is used for classification of agricultural products as an alternative to statistical and ANN approach [15]. In this study, 9 geometrical and 3 color features are obtained, and the data set is produced from digital images of 5 different wheat type (Bezostaja, Gerek, Çeşit-1252, Dağdaş, Kızıltan) for each kernel. By using these features, wheat kernels are classified with M-SVM. BPSO is used for feature selection of data set, and by using these reduced features, wheat kernels are classified with M-SVM, and obtained results are compared.

2 Material and Methods

2.1 Samples, Image Acquisition and Image Analysis

In this study 5 wheat types is used (Bezostaja, Çeşit1252, Dağdaş, Gerek, Kızıltan). For each wheat type, 80 wheat kernels which are selected by experts are used. Digital images belong to wheat samples are obtained using desktop scanner, and image belong to this operation is given in figure 1.

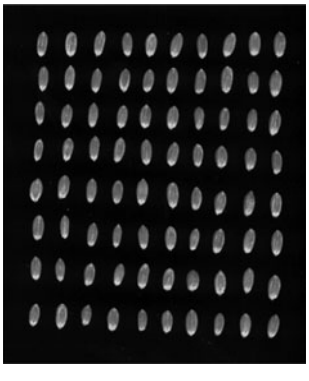


Fig. 1. Digital image belongs to wheat kernels

Color image including wheat kernels was converted into gray level image. The noise in this image was eliminated by using median filter. Gray level image was converted to black-white image via Otsu's method. A segmentation algorithm was used in order to detect wheat kernels and eliminate non-kernel points by using morphological

operations in the binary digital image. Each wheat kernel was labeled after extracting 9 geometrical features and color information from each kernel [14]. Extracted features are shown in table 1.

Table 1. Extracted features of wheat kernels

Feature	Definition
Area	Number of pixels
A	Major axis
B	Minor axis
Perimeter	$\pi.\sqrt{\frac{A^2+B^2}{2}}$
Equivalent diameter	$\frac{4.\text{Area}}{\text{Perimeter}}$
Eccentricity	$\frac{\sqrt{A^2-B^2}}{2.A}$
Roundness	$\frac{4.\text{Area}}{\pi.\text{A}^2}$
Shape factor	$\frac{4.\pi.\text{Area}}{\text{Perimeter}^2}$
Compactness	$\frac{\sqrt{\frac{4.\text{Area}}{\pi}}}{\text{A}}$
R, G, B	$\frac{1}{n}\sum_{k=1}^{n} x_k$

2.2 Support Vector Machine (SVM)

SVM is a binary classification approach which is based on statistical theory improved by Vapnik [16]. In SVM, to find the most appropriate hyperplane which can separate two classes in attribute space is aimed. According to input vector, classes are labeled as {+1,-1}. Firstly, the data which will be classified are moved to feature space by kernel function and then separated by hyperplane [17, 18]. A sample hyperplane is shown figure 2.

Optimal separating hyperplane is calculated by solving following problem:

$$\min \frac{1}{2}\sum_{i=1}^{A} w_i^2 + C\sum_{i=1}^{N} \xi_i \tag{1}$$

subject to

$$\begin{aligned} & y_i(w^T x_i + b) \geq 1 - \xi_i \forall i \in \{1,..N\} \\ & \xi_i \geq 0 \forall i \in \{1,..N\} \end{aligned} \tag{2}$$

ξ : *error parameter*

N: number of objects in training set

b: bias

C: penalty parameter

The decision function is

$$f(x) = \text{sgn}(\, w\Phi(x) + b) \tag{3}$$

Φ : *Kernel function*

SVM is a binary classification method. In multi-class classification problems, data set is separated to subsets and classification is implemented for each subset individually. Then, the results which are obtained from these binary subsets are evaluated together. In this study, according to the one-to-one strategy, binary subsets are formed, and results are evaluated using possibility calculation.

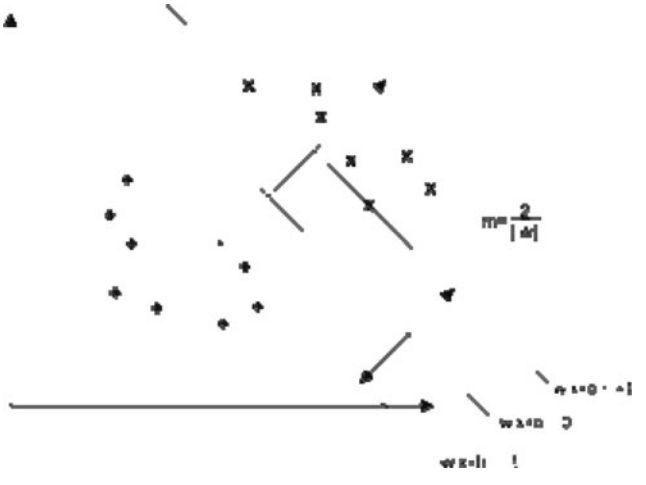


Fig. 2. Optimal separating hyper plane. *(m:maximum distance).*

2.3 Binary Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) algorithm is developed by Eberhart and Kennedy (1995). First presentation of this algorithm, like development of many algorithm used in this days, developed by modeling birds and fishes flocks movement was in 1995 [19].

In PSO technique, each potential solution is defined as particle. In a j-dimensional feature space, each particle has a velocity ($v_{i,j}$) and a position ($x_{i,j}$). Result set which is composed from particles is named as swarm. At the beginning of the algorithm, each particle has a random value in solution space. Suitability of each particle is evaluated by a fitness function. Through the iteration process, best instance of each particle is kept as best local particle ($P_{best\ i,j}$), and the best instance of the swarm is kept as best global particle ($G_{best\ i,j}$). In each iteration, velocity and position of each particle are updated using 4 and 5 [19]

$$v_{i,j}(t+1) = \omega v_{i,j}(t) + c_1 R_1 (p_{best_{i,j}} - x_{i,j}(t)) + c_2 R_2 (g_{best_{i,j}} - x_{i,j}(t)) \tag{4}$$

$$x_{i,j}(t+1) = x_{i,j}(t) + v_{i,j}(t+1) \tag{5}$$

Where,
i: Index of particle,
j: Position index of particle,
t: Number of iteration,
$v_{i,j}$: Speed parameter,
R_1,R_2: Random coefficients [0,1],
c_1,c_2: Acceleration numbers [=2],
ω: Inertial weight [=1].

In literature, both c_1 and c_2 parameters are commonly used as 2, and ω is commonly used as 1.

Binary particle swarm optimization (BPSO) firstly presented by Eberhart and Kennedy (1997) who are the developers of PSO. Different from PSO, the values belong to particle positions are not composed of real numbers but composed of binary values. In BPSO, the positions of the particles are evaluated by the velocity of the particles using sigmoid function. The positions of the particles are updated by implementing eq. 6. [20]

$$x_{i,j}(t+1) = \begin{cases} 0 & if \quad rand() >= S(v_{i,j}(t+1)) \\ 1 & if \quad rand() < S(v_{i,j}(t+1)) \end{cases} \tag{6}$$

Rand() is function which produces a random value in regular distribution between [0,1], and S is the sigmoid function given in eq. 7.

$$S(v_{i,j}(t+1)) = \frac{1}{1+e^{-v_{i,j}(t+1)}} \tag{7}$$

3 Results and Discussion

For each wheat kernel, 12 descriptive features are obtained from the digital images of wheat types. The 400x12 sized dataset is created. Each feature is normalized into [0 1]. Dataset is divided into 2 subsets using cross validation approach, and one fold is used for training and the other is used for testing. For the classification process, M-SVM is used. One-against-one strategy is selected for fusion algorithm in M-SVM. LIBSVM package [21, 22] is used for the application. Radial basis function is used for kernel function in M-SVM. This process is repeated 250 times and average results are calculated.

At first application, classification process is implemented by using whole 12 descriptive features. Firstly, grid search approach is used for kernel parameter selection $(c \in \{-15,15\}; \gamma \in \{-15,15\})$, and C and γ parameters are obtained for the optimum M-SVM model. Training and testing process are repeated 250 times. Finally average classification accuracy rate is found as 91.5% for M-SVM.

At the second application, to increase the classification accuracy, BPSO method is used in order to selecting some features in features vector. By using BPSO algorithm, some features which has negative effects or no effects in classification process are excluded from the feature vector. 5 features (area, minor axis, eccentricity, equivalent diameter and perimeter) are eliminated using BPSO algorithm. By using new feature vector (major axis, compactness, shape factor, roundness, R, G and B), average classification accuracy is increased to 92.02%.

A similar study is implemented using PCA algorithm on the same dataset [14]. Suggested BPSO approach in this study produces more successful classification results than both the other methods classification results. The comparison of the results is given in table 2.

Table 2. Classification Results (Obtained using 2-fold cross-validation and repeated 250 times)

Methods	Parameters		Accuracy (%)
	C		
M-SVM	2^{8}	2^{-2}	91.50
(12 when used feature)			
BPSO+M-SVM	2^{11}	2^{-3}	92.02
(7 when used feature)			
PCA+M-SVM	2^{9}	2^{-7}	91.08

C: cost parameter,
γ: gamma parameter in kernel function

When the results are assessed, it is shown that BPSO can be used effectively for the feature selection process in wheat variety classification. The reduction of the feature vector increases the classification accuracy and decreases the training time.

Identifying the wheat's type is important for buying and handling. Wheat which is operated in mill generally belongs to one type. This study shows that the results obtained from the proposed BPSO-M-SVM system is acceptable level in practice usage. It is seen from the results that the suggested method can be used as a reliable, accurate and fast variety recognition technique.

Acknowledgement

This study has been supported by Scientific Research Project of Selçuk University.

References

1. Du, C.-J., Sun, D.-W.: Recent developments in the applications of image processing techniques for food quality evaluation. Trends in Food Science & Technology 15, 230–249 (2004)
2. Berrueta, L.A., Alonso-Salces, R.M., Héberger, K.: Supervised pattern recognition in food analysis. Journal of Chromatography A 1158, 196–214 (2007)

3. Visen, N.S., Paliwall, J., Jayas, D.S., White, N.D.G.: Image analysis of bulk grain samples using neural Networks. Canadian Biosystems Engineering 46, 7.11–7.15 (2004)
4. Zapotoczny, P., Zielinska, M., Nita, Z.: Application of image analysis for the varietal classification of barley: Morphological features. Journal of Cereal Science 48, 104–110 (2008)
5. Choudhary, R., Paliwal, J., Jayas, D.S.: Classification of cereal grains using wavelet, morphological, colour, and textural features of non-touching kernel images. Biosystems Engineering 99, 330–337 (2008)
6. Tahir, A.R., Neethirajan, S., Jayas, D.S., Shahin, M.A., Symons, S.J., White, N.D.G.: Evaluation of the effect of moisture content on cereal grains by digital image analysis. Food Research International 40, 1140–1145 (2007)
7. Ramalingam, G., Neethirajan, S., Jayas, D.S., White, N.D.G.: Charecterization of the Influence of Moisture Content on Single Wheat Kernels Using Machine Vision. In: CSBE/SCGAB 2009 Annual Conference, Prince Edward Island, July 12-15 (2009), Paper No: CSBE09-708
8. Neethirajan, S., Karunakaran, C., Symons, S., Jayas, D.S.: Classification of vitreousness in durum wheat using soft X-rays and transmitted light images. Computers and Electronics in Agriculture 53, 71–78 (2006)
9. Wang, N., Zhang, N., Dowell, F., Pearson, T.: Determination fo durum vitreousness using transmissive and reflective images. In: 2003 ASAE Annual International Meeting, Las Vegas, Nevada, USA, July 27-30 (2003)
10. Utku, H., Köksel, H.: Use of Statistical Filters in the Classification of Wheats by Image Analysis. Joumat of Food Engineering 36, 385–394 (1998)
11. Shouche, S.P., Rastogi, R., Bhagwat, S.G., Sainis, J.K.: Shape analysis of grains of Indian wheat varieties. Computers and Electronics in Agriculture 33, 55–76 (2001)
12. Dubey, B.P., Bhagwat, S.G., Shouche, S.P., Sainis, J.K.: Potential of Artificial Neural Networks in Varietal Identification using Morphometry of Wheat Grains. Biosystems Engineering 95(1), 61–67 (2006)
13. Raudys, S., Baykan, Ö.K., Babalık, A., Denisov, V., Bielskis, A.A.: Classifiers Fusion in Recognition of Wheat Viarieties. LNCS, vol. 447, pp. 62–71 (2007)
14. Babalık, A., Baykan, Ö.K., Botsalı, F.M.: Classification of Wheat Kernels Using Multi-Class Support Vector Machine. In: ISCSE 2010, International Symposium on Computing in Science & Engineering (2010) (article in press)
15. Huang, Y., Lan, Y., Thomson, S.J., Fang, A., Hoffmann, W.C., Lacey, R.E.: Development of soft computing and applications in agricultural and biological engineering. Computers and Electronics in Agriculture 71, 107–127 (2010)
16. Vapnik, V.N.: The nature of statistical learning theory. Springer, New York (1995)
17. Avşar, E., Kamaşak, M.E., Çataltepe, Z.: Tek-sınıf destek vector makineleri kullanılarak EEG işaretlerinin sınıflandırılması. In: BIYOMUT 2009, İzmir, Turkey, May 20-24 (2009)
18. Ekici, S., Yıldırım, S., Poyraz, M.: Mesafe korumak için bir örüntü tanıma uygulaması. Gazi Üniversitesi Mühendislik – Mimarlık Fakültesi Dergisi (24), 51–61 (2009)
19. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, vol. IV, pp. 1942–1948. Piscataway, NJ (1995)
20. Eberhart, R.C.: A discrete binary version of the particle swarm algorithm. In: Proceedings of 1997 conference systems man cybernetics, pp. 4104–4108. Piscataway, NJ (1997)
21. Chang, C.C., Lin, C.J.: LIBSVM:a library for support vector machines (2001), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
22. Hsu, C.W., Lin, C.J.: A Comparasion of Methods for Multiclass Support Vector Machines. IEEE Transactions on Neural Networks 13(2), 415–425 (2002)

A Comparison of Artificial Intelligence Methods on Determining Coronary Artery Disease

İsmail Babaoğlu[1], Ömer Kaan Baykan[1], Nazif Aygül[2], Kurtuluş Özdemir[3], and Mehmet Bayrak[4]

[1] Selçuk University, Department of Computer Engineering, 42075, Konya, Turkey
[2] Selçuk University, Department of Cardiology, Selçuklu Faculty of Medicine, 42075, Konya, Turkey
[3] Selçuk University, Department of Cardiology, Meram Faculty of Medicine, 42080, Konya, Turkey
[4] Mevlana University, Department of Electrical and Electronics Engineering, 42075, Konya, Turkey
{ibabaoglu,obaykan,naygul,kozdemir}@selcuk.edu.tr, mbayrak@mevlana.edu.tr

Abstract. The aim of this study is to show a comparison of multi-layered perceptron neural network (MLPNN) and support vector machine (SVM) on determination of coronary artery disease existence upon exercise stress testing (EST) data. EST and coronary angiography were performed on 480 patients with acquiring 23 verifying features from each. The robustness of the proposed methods is examined using classification accuracy, k-fold cross-validation method and Cohen's kappa coefficient. The obtained classification accuracies are approximately 78% and 79% for MLPNN and SVM respectively. Both MLPNN and SVM methods are rather satisfactory than human-based method looking to Cohen's kappa coefficients. Besides, SVM is slightly better than MLPNN when looking to the diagnostic accuracy, average of sensitivity and specificity, and also Cohen's kappa coefficient.

Keywords: Exercise stress testing, coronary artery disease, support vector machine, artificial neural networks.

1 Introduction

Coronary artery disease (CAD) is the primary cause of mortality and morbidity in both developed and developing countries and its incidence is increasing rapidly worldwide. More than a half of the deaths in the United States are being occurred depending on cardiovascular diseases, many of which are consist of CAD [1]. Coronary angiography (CAG) is the gold standard diagnostic tool in the diagnosis of suspected CAD. However; because it is invasive and expensive method, it is not suggested as the first choice method. Exercise stress testing (EST) is a non-invasive, relatively cheap, reproducible and safe method; therefore, it can be used as one of the first choice non-invasive diagnostic tools in the diagnosis of suspected CAD. Nonetheless, the relatively low sensitivity and specificity of EST for diagnosing CAD, has led to limit its clinical usage [2,3].

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 18–26, 2010.

Recently, the artificial intelligence (AI), out of than invasive and non-invasive diagnostic tools, becomes the promising method in the diagnosis of heart diseases. Least squares support vector machine and backpropagation artificial neural network are used to classify the extracted features which are obtained from Doppler signals of the heart valve [4]. Electrocardiography (ECG) signals are classified to 10 different arrhythmias using a new fuzzy clustering neural network architecture for early diagnosis [5]. Fuzzy weighted pre-processing and artificial immune recognition system are used to classify ECG arrhythmia as a new method for the medical diagnosis [6]. An expert diagnosis system is presented for interpretation of the Doppler signals of the heart valve diseases. To make the feature extraction from Doppler signals on the time-frequency domain, wavelet transforms and short times Fourier transform methods are used. Wavelet entropy method and back-propagation neural network are employed to classify the extracted features [7]. An adaptive neuro-fuzzy network is developed to classify heart abnormalities in 10 different cardiac states and classification accuracy is more than 94% [8]. For classification of carotid artery Doppler signals in the early phase of atherosclerosis, principle component analysis and fuzzy c-means clustering methods with complex-valued artificial neural network are used [9].

More specific usage of AI methods for the determination of CAD diagnosis could also be found in literature. Lapuerta et al. investigated artificial neural network (ANN) performance to predict the occurrence of CAD based on information from serum lipid profile [10]. Süt et al. examined the diagnostic performances of multilayer perceptron neural networks (MLPNNs) for predicting coronary artery disease and compared them with different types of artificial neural network methods, namely recurrent neural networks as well as two statistical methods (quadratic discriminant analysis and logistic regression) [11]. Scott et al. searched that ANN's can determine the presence or absence of CAD with predictive accuracy equal to that of standard (expert reader clinical interpretation using imaging) clinical and stress test data [12]. Kurt et al compare performances of logistic regression, classification and regression tree, multi-layer perceptron, radial basis function and self-organizing feature maps in order to predict the presence of CAD by using demographic and medical data [13]. Zhidong proposed noninvasive diagnosis method of coronary artery disease based on the instantaneous frequency estimation of diastolic murmurs and support vector machine (SVM) classifier [14]. And also many applications carried out for diagnosing coronary artery stenosis [15-17].

In this study, it is aimed to explore the MLPNN and SVM significance on determination of CAD existence upon EST data. A proper comparison is also performed for both the MLPNN and SVM methods.

2 Materials and Method

2.1 Data Acquisition

Four hundred and eighty patients who underwent EST and CAG were included to the study. Baseline demographic characteristics, rest and peak exercise heart rate, blood pressure, exercise time were recorded. The EST results were evaluated by 2 experienced cardiologists (human-based method). ST segment depression and elevation

occurred 60 ms after the J point were recorded at each derivation in peak exercise. According to human-based method, an exercise test result was considered positive if there was ≥1 mm horizontal or downsloping ST depression or ST elevation in two contiguous leads. Within the first month following the EST, CAG was performed to all patients, and the angiographic images were evaluated by 2 skilled cardiologists. Presence of ≥50% narrowing in left main coronary artery, or ≥70% narrowing in other major epicardial coronary arteries indicated significant CAD. Patients with bundle branch blocks (right or, left bundle branch block), pre-excitation syndromes, atrial fibrillation, left ventricular hypertrophy and taking the digoxin were excluded from the study.

2.2 Multi Layered Perceptrons

A neural network generally consist of a set of neurons, a pattern of connectivity, a propagation rule, an activation rule, a transfer function and a learning rule [18]. Artificial neural networks can be design many architectures and structures. Multi-layered Perceptrons (MLPs) are simple and most frequently used ANN architectures [19]. MLP model is a feed-forward network and as shown Fig. 1 which consists of one input layer, one or more hidden layers and one output layer. Layers can have different number of neurons. The input signals x_i are dispatched to neurons in the hidden layer by using input layer neurons. Each neuron at hidden layers or output layer receives a weighted sum from all neurons in the previous layer [20]. Outputs of the hidden layer neurons together with the output layer neurons are calculated with defined transfer functions (f). Neuron outputs are calculated as:

$$y_j = f\left(\sum w_{ji} x_i\right) \tag{1}$$

here f can be a sigmoidal or a hyperbolic tangent function, w_{ji} is the weight. For setting the weights of ANN, many learning algorithms can be adopted. In this study, MLP neural networks are trained with backpropagation (BP) learning algorithm. In BP algorithm, error (E) calculated as the sum of squared differences between the desired and actual values of the output neurons which propagate through the layers of neurons to update the weights. E is defined in Eq.(2) below as;

$$E = \frac{1}{2}\sum_j \left(y_{dj} - y_j\right)^2 \tag{2}$$

where, y_j is the actual value of output neuron and y_{dj} is the desired value of that neuron [20,21].

2.3 Support Vector Machines

Support vector machine is proposed by Vapnik (1995) based on structural risk minimization (SRM) principle. SVM as a new machine learning technique is used for many purpose such as classification, recognition, regression [22]. SVMs can classify

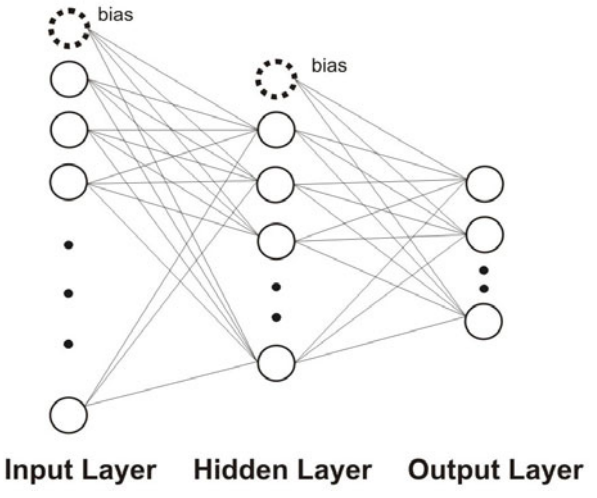


Fig. 1. A Multi Layered Perceptron

the input samples when the classes are linearly separable. If input space is nonlinear, SVMs need mapping N dimensional input space to a high-dimensional feature space using several kernel functions such as polynomial kernel, dot product, radial basis function kernel [23]. SVMs search an optimal separating hyper-plane that maximizes the margin between itself and the nearest training examples in the new high-dimensional space. A separating hyper-plane is a linear function and it can divide the training data into two classes. The training examples that are closest to the hyper-plane are called as support vectors. SVMs can be used both binary classifications and multi-classes problems [24-27].

Commonly used kernel functions can be briefly given as:

Dot product kernels : $K\left(x, x^{'}\right) = x.x^{'}$

Polynomial kernels : $K\left(x, x^{'}\right) = \left(x.x^{'} + 1\right)^{d}$; where d is the degree of kernel.

RBF kernels : $K\left(x, x^{'}\right) = \exp\left(-\left\|x - x^{'}\right\|^{2} / \sigma^{2}\right)$; where σ is positive real number.

In this study, prediction of CAD problem is considered a binary classification problem and SVMs are also applied to solve it. Four kernel functions namely linear, polynomial, radial basis, and sigmoid are tested and the best one is adopted.

2.4 K-Fold Cross Validation

To make the test results more meaningful and benefitable, *k*-fold cross validation method which minimizes the bias association with the random sampling of the training can be used [28,29]. Whole data is randomly divided to *k* mutually exclusive and approximately equal size subsets. Training and test processes are performed *k* times.

In each case, one of the folds is taken as test data and the remaining folds are added to form training data. So *k* times different test results are obtained. The average of these results gives the test accuracy of the algorithm.

2.5 Screening Test

Performance evaluations of proposed methods are implemented using screening test in point of sensitivity, specificity, positive and negative predictivity and accuracy. In this test TP, FP, TN and FN as described as follow [30];

True Positives (TP) : Those who test positive for a condition and are positive
False Positives (FP) : Those who test positive, but are negative.
True Negatives (TN) : Those who test negative and are negative.
False Negatives (FN) : Those who test negative but are positive.

Positive Predictive Value (PPV): Percent of patients with positive test having disease. PPV is calculated as;

$$PPV = \frac{TP}{TP + FP} \tag{3}$$

Negative Predictive Value (NPV): Percent of patients with negative test that do not have disease. NPV is calculated as;

$$NPV = \frac{TN}{TN + FN} \tag{4}$$

Sensitivity (SEN) and Specificity (SPE): Independent of disease prevalence in the community. SEN, SPE and Diagnostic Accuracy (ACC) are calculated as [6, 30];

$$SEN = \frac{TP}{TP + FN} \tag{5}$$

$$SPE = \frac{TN}{FP + TN} \tag{6}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

2.6 Cohen's Kappa Coefficient

Cohen's kappa coefficient (κ) is a statistical measure of inter-rater reliability and measures the agreement between the evaluations of two raters when both are rating the same object. A kappa coefficient of 1 indicates the perfect agreement. A kappa coefficient of 0 indicates the agreement is no better than chance. However by no means universally accepted the ranges of κ mentioned on the previous studies are $\kappa < 0$ no agreement, $0 \leq \kappa \leq 0.20$ very low agreement, $0.21 \leq \kappa \leq 0.40$ low agreement, $0.41 \leq \kappa \leq 0.60$ moderate agreement, $0.61 \leq \kappa \leq 0.80$ full agreement, $0.81 \leq \kappa \leq 1.00$ almost perfect agreement [31,32].

In our study, we use Cohen's kappa coefficient to compare specifically the three methods (human-based, SVM and MLPNN) in order to outline the rate of agreement with CAG results. Cohen's kappa coefficients together with the test results are shown in Table 3.

3 Results

In this study, on determination of coronary artery disease existence upon EST data, 3 different methods are used. Data set consist of 480 patient and 23 verifying features. 346 of these patients have CAD and rest of 134 is healthy.

Each feature used in the study is normalized into [-1 1] range. Training and test implementations are applied by using k-fold cross validation method being k as 5. In this situation, for each fold, size of training data set is 384*23 and size of test data set is 96*23.

Human-based method can be expressed as the evaluation of the EST results by 2 experienced cardiologists. On the other hand, different MLP and SVM architectures are trained and tested in the study. Grid search algorithm is used on these training and test processes for both MLPNN and SVM techniques [33]. In grid search algorithm, the values of each parameter across the specified search range is tried to find optimum ones using geometric steps. The value ranges of the parameters used by the grid search algorithm for both different MLPNN and SVM models and are given in Table 1 and Table 2 respectively. The highest value of the difference between the diagnostic accuracy and the training error is selected as the optimum classification model. In other words, MLPNN and SVM models are attempted to minimize the training error while maximizing the diagnostic accuracy.

According to the MLPNN evaluation, tangent sigmoid transfer function is used at the hidden layer and logarithmic sigmoid transfer function is used at the output layer in the best result. For this network, hidden layer has got 75 neurons. Average test accuracy and training errors are found to be 78.13% and 1.86%.

In SVM evaluation, the best results are obtained by using radial basis kernel. In this best SVM model, γ value is 0.4 and C value is 10. Average test accuracy and training errors are found to be 79.17% and 1.30%.

Results obtained with all 3 methods are given in Table 3. The results obtained from different MLP and SVM methods are examined to select the optimum models for each method. While selecting the optimum models, the primary aim is to select higher

Table 1. MLPNN parameters search grid

Prm	SVal	EVal	Int
lr	0.1	0.5	0.1
mc	0.5	0.9	0.1
neuron	5	200	5
iteration	1000	10000	1000

Prm, Name of the parameter; SVal, Start value of the parameter; EVal, End value of the parameter; Int, Interval of the parameter between start value and end value; lr, Learning rate; mc, Momentum coefficient; neuron, Number of the neurons in the hidden layer; iteration, Training iterations size.

Table 2. SVM parameters search grid

Kernel	Prm	SVal	EVal	Int
Linear	c	1	10	1
Polynomial	d	1	5	1
Polynomial	γ	0	1	0.1
Polynomial	r	0	5	1
Polynomial	c	1	100	1
RBF	γ	0	5	0.1
RBF	c	1	100	1

Kernel, Kernel type used in SVM; Prm, Name of the parameter; SVal, Start value of the parameter; EVal, End value of the parameter; Int, Interval of the parameter between start value and end value; Linear, Linear type kernel; Polynomial, Polynomial type kernel; RBF, Radial basis type kernel; c, Parameter of linear kernel; d, γ, r and c, Parameters of polynomial kernel; γ and c, Parameters of RBF kernel.

diagnostic accuracy and sensitivity, the secondary aim is to select the lower training error. After optimum model selection for both methods, Cohen's kappa coefficients are calculated. Obtained values by human-based method were consistent with that of obtained in the literature. Nonetheless, sensitivity (mean 72%, range 45–92%) and specificity (mean 77%, range 17–92%) values are relatively variable in the diagnosis of CAD using EST by human-based evaluation [1-3]. Disparity in positive criterion of EST, prevalence/intensity of CAD and interobserver variability could be effective factors in this variability. In our study, a human-based interpretation of EST showed relatively high sensitivity of 78% but a poor specificity of 43% and poor agreement with CAG results (κ 0.21). In contrast, both SVM and MLPNN methods provided a better sensitivity, specificity and diagnostic accuracy and as expected NPV and PPV, compared with human-based method. It is also observed that both the two methods agreement with coronary angiographic results was rather satisfactory. Besides, SVM is slightly better than MLPNN when looking to the diagnostic accuracy, average of sensitivity and specificity, and also kappa coefficient. In terms of explication, well education is needed to improve the usage of EST. Therefore, the supporting of human-based method with AI methods will provide standardization, and remove disparity of personal explication.

Table 3. Test Results

Method	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	κ
Human-based	68.33	78.03	43.28	78.03	43.28	0.213
SVM	79.17	84.76	63.98	86.70	59.69	0.473
MLPNN	78.13	86.35	59.64	83.42	60.23	0.411

ACC, Diagnostic accuracy; SEN, Sensitivity; SPE, Specificity; PPV, Positive predictive value; NPV, Negative predictive value; SVM, Support vector machine; MLPNN, Multi layered perceptron neural network.; κ, Cohen's kappa coefficient.

4 Discussion

The findings of our study suggest that assessment of EST using AI methods increase sensitivity, specificity and diagnostic accuracy compared to the assessment of EST by

human-based method. Improvement in the sensitivity and specificity of EST for diagnosis of CAD may improve the PPV and NPV of the test in population with suspected CAD. This situation enhances the usage of EST as a more reliable test which is still commonly used in clinical practice.

Acknowledgement

This study has been supported by Scientific Research Project of Selçuk University.

References

1. Gibbons, R., Balady, G., Bricker, J.T., et al.: ACC/AHA Guideline update for exercise testing: summary article. A report of the American college of cardiology/American heart association task force on practice guidelines (Committee to update the 1997 exercise testing guidelines). J. Am. Coll. Cardiol. 40(8), 1531–1540 (2002)
2. San Roman, J.A., Vilacosta, I., Castillo, J.A., Rollan, M.J., Hernandez, M., Peral, V., Garcimartin, I., del Mar de la Torre, M., Fernandez-Aviles, F.: Selection of the optimal stress test for the diagnosis of coronary artery disease. Heart 80(4), 370–376 (1998)
3. Thom, T., Haase, N., Rosamond, W., Howard, V.J., Rumsfeld, J., Manolio, T., Zheng, Z.J., Flegal, K., O'Donnell, C., Kittner, S., Lloyd-Jones, D., Goff Jr., D.C., Hong, Y., Adams, R., Friday, G., Furie, K., Gorelick, P., Kissela, B., Marler, J., Meigs, J., Roger, V., Sidney, S., Sorlie, P., Steinberger, J., Wasserthiel-Smoller, S., Wilson, M., Wolf, P.: Heart disease and stroke statistics–2006 update: A report from the American heart association statistics committee and Stroke statistics subcommittee. Circulation 113, e85–e151 (2006)
4. Comak, E., Arslan, A., Turkoglu, I.: A decision support system based on support vector machines for diagnosis of the heart valve diseases. Computers in Biology and Medicine 37, 21–27 (2007)
5. Ozbay, Y., Ceylan, R., Karlik, B.: A fuzzy clustering neural network architecture for classification of ECG arrhythmias. Computers in Biology and Medicine 36(4), 376–388 (2006)
6. Polat, K., Sahan, S., Gunes, S.: A new method to medical diagnosis: artificial immune recognition system (AIRS) with fuzzy weighted pre-processing and application to ECG arrhythmia. Expert Systems with Applications 31(2), 264–269 (2006)
7. Turkoglu, I., Arslan, A., Ilkay, E.: An expert system for diagnosis of the heart valve diseases. Expert Systems with Applications 23(3), 229–236 (2002)
8. Kannathal, N., Lim, C.M., Rajendra Acharya, U., Sadasivan, P.K.: Cardiac state diagnosis using adaptive neuro-fuzzy technique. Medical Engineering & Physics 28(8), 809–815 (2006)
9. Ceylan, M., Ceylan, R., Dirgenali, F., Kara, S., Ozbay, Y.: Classification of carotid artery doppler signals in the early phase of atherosclerosis using complex-valued artificial neural network. Computers in Biology and Medicine 37, 28–36 (2007)
10. Lapuerta, P., Azen, S.P., Labree, L.: Use of neural networks in predicting the risk of coronary artery disease. Computers and Biomedical Research 28(1), 38–52 (1995)
11. Süt, N., Şenocak, M.: Assessment of the performances of multilayer perceptron neural networks in comparison with recurrent neural networks and two statistical methods for diagnosing coronary artery disease. Expert Systems 24(3), 131–142 (2007)
12. Scott, J.A., Aziz, K., Yasuda, T., Gewirtz, H.: Integration of clinical and imaging data to predict the presence of coronary artery disease using neural networks. Journal of Nuclear Cardiology 11(4), 26 (2004)

13. Kurt, I., Ture, M., Kurum, A.T.: Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Systems with Applications 34, 366–374 (2008)
14. Zhidong, Z.: Noninvasive diagnosis of coronary artery disease based on instantaneous frequency of diastolic murmurs and SVM. In: Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China, pp. 5651–5654 (2005)
15. Cios, K.J., Goodenday, L.S.: Hybrid intelligence system for diagnosing coronary stenosis. Combining fuzzy generalized operators with decision rules generated by machine learning algorithms. IEEE Engineering in Medicine and Biology Magazine 13(5), 723–729 (1994)
16. Mobley, B.A., Moore, W.E., Schechter, E., Eichner, J.E., McKee, P.A.: Neural network predictions of significant coronary artery stenosis in women. Computational Intelligence Processing in Medical Diagnosis, 261–290 (2002)
17. Mobley, B.A., Moore, W.E., Schechter, E., Eichner, J.E., McKee, P.A.: Neural network predictions of significant coronary artery stenosis in men. Artificial Intelligence in Medicine 34, 151–161 (2005)
18. Rumelhart, D.E., McClelland, J.L.: PDP Research Group: Paralel distributed processing: exploration in the microstructure of cognition, vol. 1, p. 950. MIT Press, MA (1989)
19. Sagiroglu, S., Colak, I., Bayindir, R.: Power factor correction technique based on artificial neural networks. Energy Conversion and Management 47(18-19), 3204–3215 (2006)
20. Serhatlioglu, S., Hardalac, F., Guler, I.: Classification of transcranial doppler signals using artificial neural network. Journal of Medical Systems 27, 205–214 (2003)
21. Pao, Y.H.: Adaptive pattern recognition and neural networks. Addison-Wesley, Reading (1989)
22. Vapnik, V.: The nature of statistical learning theory. Springer, New York (1995)
23. Gunn, S.R.: Support vector machines for classification and regression. ISIS, Technical Report, University of Southampton, Department of Electrical and Computer Science (1998)
24. Chen, K.Y., Wang, C.H.: A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan. Expert Systems with Applications 32(1), 254–264 (2007)
25. Kulkarni, A., Jayaraman, V.K., Kulkarni, B.D.: Support vector classification with parameter tuning assisted by agent-based technique. Computers and Chemical Engineering 28, 311–318 (2004)
26. Seo, K.K.: An application of one-class support vector machines in content-based image retrieval. Expert Systems with Applications 33(2), 491–498 (2007)
27. Takeuchi, K., Collier, N.: Bio-medical entity extraction using support vector machines. Artificial Intelligence in Medicine 33(2), 125–137 (2003)
28. An, S., Liu, W., Venkatesh, S.: Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression. Pattern Recognition 40, 2154–2162 (2007)
29. Shao, J.: Linear model selection by cross-validation. Journal of American Statistical Association 88, 486–494 (1993)
30. Nielsen, C., Lang, R.S.: Principles of screening. The Medical Clinics of North America 83(6), 1323–1337 (1999)
31. Agresti, A.: An introduction to categorical data analysis, 2nd edn. Wiley, Chichester (2007)
32. Sprent, P., Smeeton, N.C.: Applied non-parametric statistical methods, 3rd edn. Chapman & Hall, Boca Raton (1999)
33. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to algorithms, 2nd edn. MIT Press, Cambridge (2001)

Using Chaotic System in Encryption

Oğuz Findik and Şirzat Kahramanli

Selçuk University, Department of Computer Engineering, 42075, Konya, Turkey
{oguzf,skahramanli}@selcuk.edu.tr

Abstract. In this paper chaotic systems and RSA encryption algorithm are combined in order to develop an encryption algorithm which accomplishes the modern standards. E.Lorenz's weather forecast' equations which are used to simulate non-linear systems are utilized to create chaotic map. This equation can be used to generate random numbers. In order to achieve up-to-date standards and use online and offline status, a new encryption technique that combines chaotic systems and RSA encryption algorithm has been developed. The combination of RSA algorithm and chaotic systems makes encryption system.

Keywords: Asymmetric encryption, Chaotic encryption, RSA.

1 Introduction

With the increase on widespread use of internet in nowadays, some security issue become more frequent. The main reason of this issue is that internet is an easy accessible system, and it allows the roaming data available for unwanted users [1].

The data packages are transferred from one point to another through the public available network. Thus, the packages are unsecure and accessible. This situation becomes a concerning event when dealing with some secure data. As long as the protection of this data is not possible, the internet and digital environments couldn't be accepted as secure [2].

Information security is provided by cryptology mainly in order to avoid some threats like intrusion and modification of the data and fake identification.

Cryptography, which is implemented by mathematical methods, is a technology that provides the information security, reliability, integrity and identity verification.

Cryptography deals with the following primary aims [3,11,10];

- *Confidence*: The information could not be recognized by unwanted users.
- *Integrity*: The information could not be replaced during the transfer process.
- *Undeniability*: The information could not be denied by the sender.
- *Identification*: The sender and the receiver could be identified by each other.

In this paper chaotic systems and RSA encryption algorithm are combined in order to develop an encryption algorithm which accomplishes the modern standards. In the application part of the work, the encryption implementation was developed to satisfy the user needs. Public or private key can be given in the generated program as a parameter. The application was developed by using Delphi 7.0 enterprise edition.

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 27–33, 2010.

2 Materials and Methods

2.1 Asymmetric Encryption

Asymmetric encryption method uses two different key. The usage of two different key has important advantages on reliability, key distribution and validation topics. These two different keys are named as public and private key [10].

The key which is used for encryption is different from the key which is used for decryption for all users in asymmetric encryption method. In other words, public and private keys are specific for the users. Besides, in asymmetric encryption method, private key could not be obtained from the public key.

This kind of encryption is also called as public key crypto system; because, the public key is shared to anybody. A foreigner could use the public key to encrypt some information, but only one decrypts the information which holds the private key corresponding to the public key [4].

2.2 RSA

RSA is found in 1978, and it is the abbreviation of the inventors surnames initial letters; Ronald Rivest, Adi Shamir, Leonard Adleman [5]. It is the most commonly used asymmetric encryption algorithm [6]. RSA is used in both encryption and digital signature systems. The algorithm uses factorization principle in math, and it is summarized as follows [7];

- Public and private key are generated.
- Two different prime numbers are found, P and Q.
- Two different integers N and Z are calculated as: $N=P*Q$; $Z=(P-1)*(Q-1)$;
- A new integer E which has 1 as the common divisor with Z is found
- Public key is determined as $[E,N]$
- A new integer D which obeys the rule $D=E^{-1} \bmod Z$ is found
- Private key is determined as $[D,N]$

The encryption and decryption processes are implemented using the obtained Public and Private key as follows;

If m is the message considered to be encrypted, the message is split into k-bit parts obeying the form $2^k<N$. Then, the parts of the message [m=m(1)+m(2)+m(3)+ ...+m(n)] are encrypted as $C(i)=m(i)^E \bmod N$.

After the encryption process, the encrypted message is decrypted using the private key $[D,N]$ as $m(i)=C(i)^D \bmod N$ [5].

2.3 Chaos Theory

Chaos theory is a field of study in mathematics, physics, economics and philosophy studying the behavior of dynamical systems that are highly sensitive to initial conditions. This sensitivity is popularly referred to as the butterfly effect. Small differences in initial conditions (such as those due to rounding errors in numerical computation) yield widely diverging outcomes for chaotic systems, rendering long-term prediction impossible in general.

Edward Lorenz's, expert of meteorology in Massachusetts Technology Institution [8,12], researches about atmospherically convection fact is as below: Since sun rays

heats the earth's surface and this heat reflects to air the lower layer air becomes hotter and lighter. While the hot and light air rises, cold and heavy air which is in top layer moves to below. This bidirectional movement is called as atmospherically convection. Lorenz [12] simulates this air movement which comes true in atmosphere in his own laboratory by placing thermometer to closed boxes top and bottom side. Although Lorenz [12] used two equations in this method, 3 equations in below gives the best approach according to true value.

$$dx / dt = \mu (y - x) \tag{1}$$

$$dy / dt = rx - y - xz \tag{2}$$

$$dz / dt = xy + bz \tag{3}$$

Weather forecast is identified by the alteration of x, y, z parameters given above which symbolizes the activities used in three differential equations.

X: It is proportional with the velocity of the movement of air formed by convection.
Y: It is proportional with the difference of rise and fall air temperatures.
Z: It is proportional with the difference between air's vertical top and bottom temperatures which exposed to convection

There is 3 different constant value used in equations. These are:

μ: It symbolizes ratio of air's thermal conductivity and viscosity. Generally, this value is used as 10. It is named as *Prandtl number*.
b : It symbolizes the ratio of areas width and height in which air convection is formed. For this constant Lorenz [12] has commonly used 2.6667.
r : It symbolizes the difference between systems top and bottom temperature. It is named as *Rayleigh number*. Generally, this value is used as 20 [9].

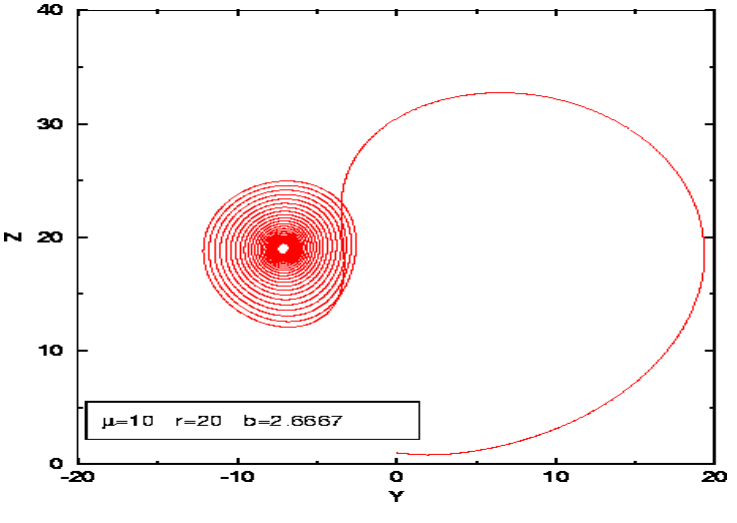


Fig. 1. Convoluted figure for Y and Z; Y=-7.11, Z=18.99 edges are formed during the iteration process

X, Y, Z variables are the initial values necessary for weather forecasting. In the sample chart given below (Fig.1), the alteration of Y depending on Z is shown by using μ=10, b=2.6667 and r=20 constant values, and X=1, Y=0, Z=1 is used as initial values.

Fig. 2 given below is gained by using μ=10, b=2.6667 and r=28 constant values

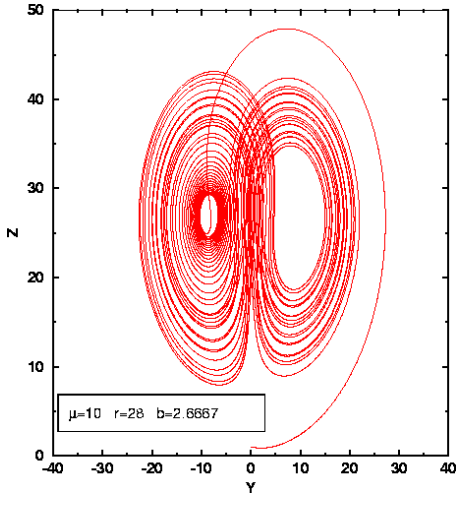


Fig. 2. Convoluted figure for Y and Z; Initial constants are same with Fig. 1 except r is used as 28

Fig. 3 gives the alteration depending on the parameters used in Fig. 2.

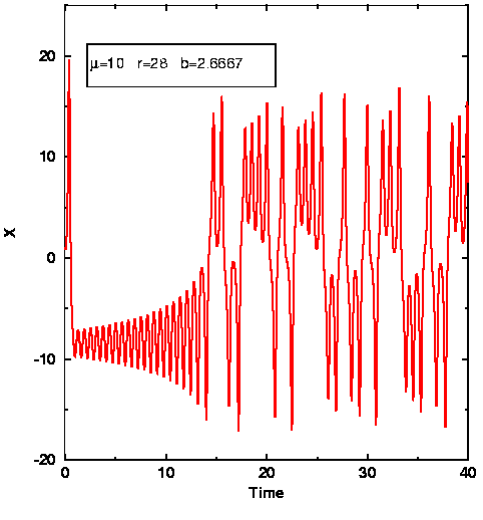


Fig. 3. The alteration of X depending on time

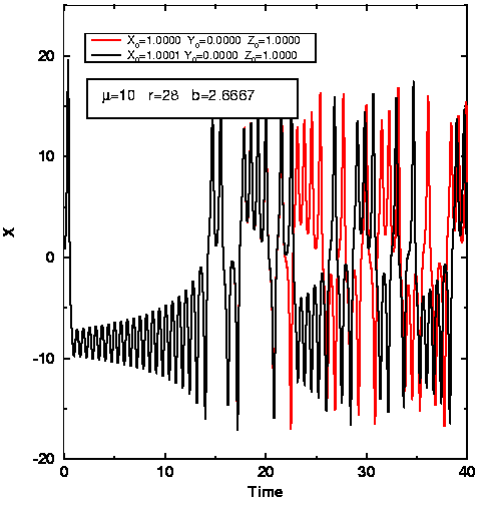


Fig. 4. The alteration of X depending on the initial values of X; X is used as 1 and 1.0001

Fig. 4 shows the alteration of convolution with a minimal change on parameter X. X is used as 1 and 1.0001.

The results show that chaotic problems solutions are depend on initial variables and initial constants tightly and sensitively. A small change in initial variables and constants causes large changes at the end.

2.4 Proposed Algorithm

In chaotic encryption system, key value depends on sensitively to the initial values. Breaking an encrypted file requires gaining of the key value. But this process requires sensitively prediction of initial values which are used in key production. Even a small round off error in key values used in decryption of the file will cause production of totally a new key, and because of this, file will not be able to be decrypted. This conclusion proves power of the chaotic data encryption method.

In this study, initial values of chaotic model which is used for encryption are added to the end of the encrypted file, and these codes are encrypted with RSA encryption algorithm. This process, provides hiding the initial values of chaotic model, and so that, by encrypting these initial values by RSA method, identification and undeniability aims are proved. The method implemented in this study can be used in digital signature applications [10].

In this study, Lorenz's differential equations are used for the production of keys. By using Lorenz's differential equations, a new number is produced subject to the initial values at the end of all triple iteration. The block size is generated as much as it will be used in encryption process, and this block is composed of new numbers generated in each iteration step [10].

In this study exclusive or (XOR) operation is used for file encrypting. According to the block sizes, XOR operation is performed to the whole file. Chaotic encrypting is too sensitive to the initial values as known, so that the initial values should be

protected safely. In the proposed method, initial values are encrypted by RSA method and added to the end of the file. Even if this encrypted file is obtained, initial values can not be extracted without the private key of RSA algorithm.

3 Experimental Result

This study shows a safe way to protect and transfer data in both online and offline states. The primary aim is constructing a novel, safe and fast encryption algorithm. The proposed algorithm performs all confidence, integrity, undeniability and identification tasks mentioned above.

A novel encryption method using RSA combining with chaotic model is introduced. Confidence and integrity tasks are ensured with chaotic blocks, and undeniability and identification tasks are ensured with RSA encryption algorithm.

Encryption algorithms blocks which used in data encryption should be formed according to some specific and random rules than constant values [10].

In order to decrypt the encrypted data, same iterations are implemented depending on the initial values after the data encrypted. In fig. 5, first few lines of the file and encrypted file are given.

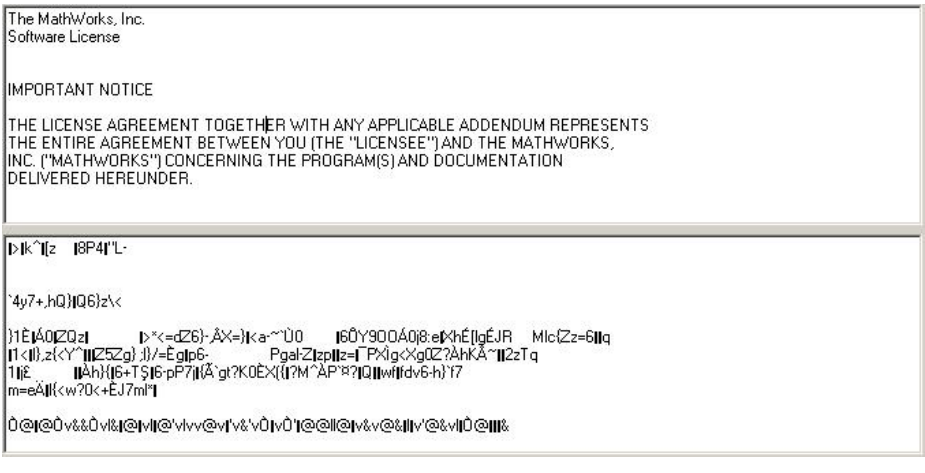


Fig. 5. 'licence.txt' file (top) and encrypted file with the proposed algorithm (bottom)

4 Conclusion

In this paper chaotic systems and RSA encryption algorithm are combined in order to develop an encryption algorithm which accomplishes the modern standards. In this study confidence and integrity are provided by using chaotic block encrypting technique and undeniability and identification are provided by RSA encrypting method.

Acknowledgement

This study has been supported by Scientific Research Project of Selçuk University.

References

1. Dalton, C.I., Griffin, J.F.: Applying military grade security to the Internet. Elsevier Computer Networks and ISDN Systems 29, 1799–1808 (1997)
2. Molva, R.: Internet security architecture. Computer Networks 31, 787–804 (1999)
3. Pasquinucci, A.: Web voting, security and cryptography. Computer Fraud & Security 3, 5–8 (2007)
4. Ahmad, D.R.M., Dubrawsky, I., et al.: Cryptography, Hack Proofing Your Network, 2nd edn., pp. 165–203 (2002)
5. Ning-bo, M., Yu-pu, H., Hai-wen, O.: Broadcast encryption schemes based on RSA. The Journal of China Universities of Posts and Telecommunications 16(1), 69–75 (2009)
6. Batinaa, L., Ors, S.B., Preneel, B., Vandewalle, J.: Hardware architectures for public key cryptography. Integration, the VLSI Journal 34, 1–64 (2003)
7. Herson, D.: The Changing Face of International Cryptography Policy: Part 15 – Trusted Third Parties. Computer Fraud & Security 11(1), 6–7 (2000)
8. Gleick, J.: Kaos. Tubitak Press (1995)
9. Lorenz, E.: J. Atmos S.20, 130 (1963)
10. Fındık, O.: Use Of Chaotic Syste. In: Cryptography, Master thesis, Selcuk University, 86 p (2004)
11. Kodaz, H.: Cryptography in data communication for security, Master thesis, Selcuk University (2002)
12. Lorenz, E.N.: The essence of chaos, University of Washington Press, Washington (1993)

Secure Medical Diagnosis Using Rule Based Mining

M.A. Saleem Durai and N.Ch. Sriman Narayana Iyengar

School of Computing Science and Engineering, VIT University, Vellore-632014, TN, India
masaleemdurai@vit.ac.in, nchsniyengar48@gmail.com

Abstract. Security is the governing dynamics of all walks of life. Here we propose a secured medical diagnosis system. Certain specific rules are specified implicitly by the designer of the expert system and then symptoms for the diseases are obtained from the users and by using the pre defined confidence and support values we extract a threshold value which is used to conclude on a particular disease and the stage using Rule Mining. "THINK" CAPTCHA mechanism is used to distinguish between the human and the robots thereby eliminating the robots and preventing them from creating fake accounts and spam's. A novel image encryption mechanism is designed using genetic algorithm to encrypt the medical images thereby storing and sending the image data in a secured manner.

Keywords: Rule Mining, Encryption, Genetic Algorithm and THINK.

1 Introduction

Diseases are the worst things to none in this earth. In diagnosing diseases, doctors have to cope with many difficulties, the patient's symptoms are usually not clear; the similarities in some disease's symptoms are difficult to distinguish. Doctors always have to test many times before making a decision. So the diagnosis result depends on not only patient's symptoms but also the doctor's experiences. Wrong decision means wrong treatment and the patient would suffer more.

The disease is determined by using a rule base, populated by rules made for different types of diseases. The patient is required to enter the symptoms and before which need to present any current medical condition if he/she posses. Both these results (symptoms and general medical conditions) help the diagnostic logic to determine the treatment for the patient with accuracy. Our diagnosis does a complex analysis of all the information gathered about our symptoms.

The main focus for the development of the system is on the architecture and the algorithm used to find the probable disease, stage and the appropriate treatment for the disease a patient may have. The objective of this work is to serve people in a more and efficient and an economic manner and as well give a secured diagnostic system where personal details of one is not accessed by anyone thereby maintaining the integrity of the system. The paper is organized as follows with the literature following this section, followed by our proposal, implementation and results, conclusions and future works followed by references.

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 34–42, 2010.

2 Related Works

In this section, we summarize the related work to our proposed research. The present day medical diagnosis system work based on fuzzy rules where symptoms from a user is taken as inputs. [1] The disease is determined by using a rule base, populated by rules made for different types of diseases. The algorithm uses the output of the rule base (i.e. the disease name) and the symptoms entered by the user to determine the stage of cancer the patient is in. Both these results (disease name and stage) help the diagnostic logic to determine the treatment for the patient with accuracy. Other work deals with the presentation of the benefits of using Data Mining techniques in the computer-aided diagnosis (CAD),[2] focusing on the cancer detection, in order to help doctors to make optimal decisions quickly and accurately. Other systems deal with medical diagnosis by framing rules and then using genetic algorithm in order to make the diagnosis in an optimized and efficient way. [3] Most of the medical diagnosis systems though are efficient do not offer the required security and hence we have focused on developing a secured medical diagnosis system.

3 Proposed Work

3.1 System Architecture

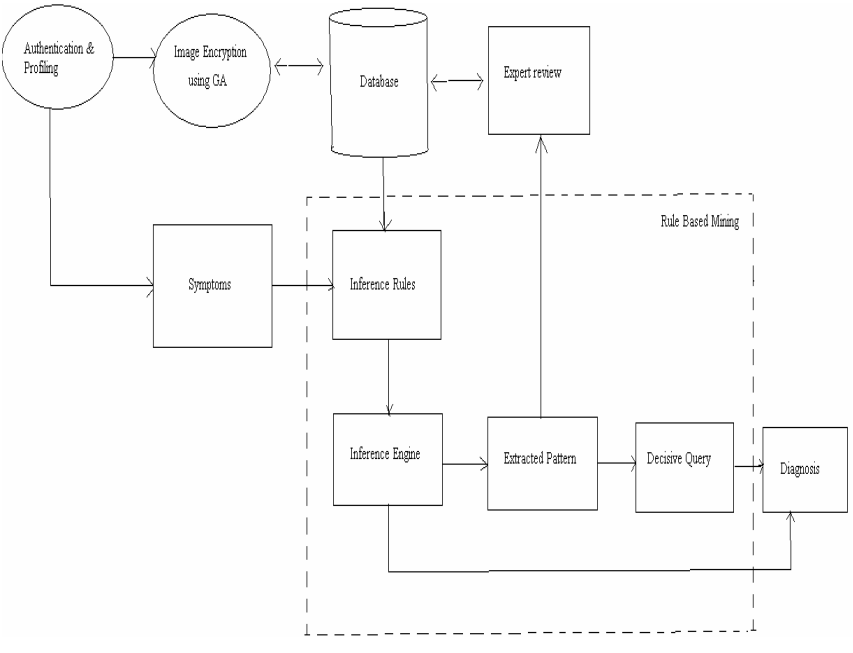


Fig. 1. Architecture of the medical diagnosis system

3.2 Procedure Used in the Proposed System

1. The users on entering the web site need to answer a THINK CAPTCHA.
2. Once they solve it. They need to give in their personal details to create their own profiles and as well select their tracking id's. In case if the user selects a tracking id that is already used, an error is reported and a new tracking id is selected by the user.
3. Once the user creates a profile, he gets to enter the symptoms in the symptoms drop down option provided.
4. Users can as well upload medically related images and reports, which are encrypted using a specially designed image encryption algorithm that functions based on the concepts of genetic algorithm.
5. Certain inference IF-THEN rules are pre specified that are used, to match the optimized input with the pre stored symptoms.
6. The support and confident values specify the minimum number of symptoms that need to be matched in order to fire a rule and make the diagnosis.
7. Nausea and Fever levels act as a threshold value in calculating the stage and the progress of a disease.
8. If symptoms match, even more than one disease might be reported which are called the pattern of disease.
9. User querying is used to mine the data from the pattern where the most useful pattern is identified and the diagnosis is made and the prognosis is followed.
10. In case if the user querying does not make a diagnosis, the case is moved on for expert review where it is reviewed by an expert and diagnosis is made.

3.3 Optimization of Input

The symptoms which are entered in the form of symptoms are optimized at the input stage itself. A drop down list is used. Once a certain symptom is selected in the first symptom drop down the rest of the symptoms are altered automatically so as to allow the users to enter only possible symptoms for a possible diseases which is predicted based on the first symptom. This prevents the users to select symptoms that may not be relevant to their medical condition and hence an optimization is performed in order to obtain efficient processing and running time.

When the symptoms are selected from the symptoms window, the individual symptom fields are stored in the database and are matched with the predefined symptoms for diseases stored in the database. If and only three symptoms out of the five symptoms match with the symptoms in the database, a particular disease is fired as the output values. Here three symptoms denote the support and confident values that needs to be met for an inference rule to be fired. Once these values are met, the inference rules fires the disease and in case if more than one disease has common symptoms, all the possible diseases are listed. This list forms the pattern of diseases from which the exact symptom is queried. In case if no other disease share common

symptom, the disease is reported and considering the allergies and medical condition of the patient, the cure is displayed.

A sample rule considering the nausea and fever levels to decide stage and diagnose a particular disease is shown below.

IF[(headache=high) AND (seizures=high) AND (memory loss =high) AND (loss of vision =high) AND (nausea=high) AND (fever=high)] THEN Brain Tumor.

IF [(nausea=high) AND (vomiting=no OR vomiting=high), OR (headache=no OR headache=high), OR (restlessness=no OR restlessness=high)] THEN Stage-2 of Tumor.

The snapshot of SMD is shown below

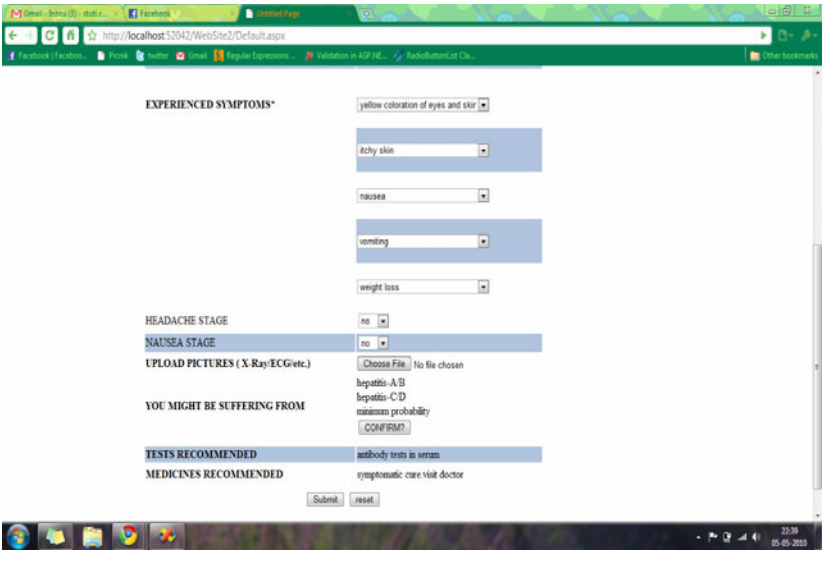


Fig. 2. Snapshot of Medical Diagnosis System

3.4 Stage Calculation

Along with the symptoms, the nausea and fever levels are also taken as inputs. Based on the levels of fever and nausea being high, low or no, we decide on the patient's stage and progress of disease.

3.5 User Querying

In case if the symptoms match, a pattern of diseases is extracted and more than one disease with the common symptom is reported. In such a case, user needs to click on the confirm button where users are queried with a key symptom and based on the answer for the query final diagnosis is made and one disease is concluded upon.

The snap shot of the user querying is shown below.

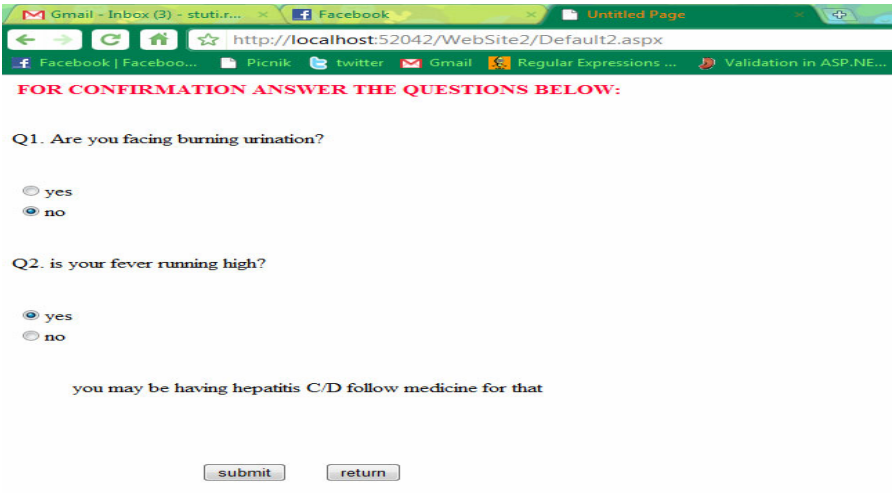


Fig. 3. Snapshot of User Querying

3.6 Image Encryption Using Genetic Algorithm

An image is considered, since the pixels in it can be viewed as a matrix, with rows and columns in an array. First the preliminary step is applied which is row-column transposition of the pixels. The inputs are automatically generated in the form of a mathematical operation say multiples of 2 and multiples of 2. Here all the multiple numbers of 2 and 2 are chosen as inputs in the form of co-ordinates with multiples of 2 taking the 'a' position and multiples of 2 taking the 'b' position. Thus corresponding output values are obtained after performing the function. Here input and output refer to the position of a pixel. Thus positions of two pixels are obtained. When the input and output pixel position values are obtained, considering the pixels to be 8 bit pixels, the 8 bits are actually divided into two parts with each part having 4 pixels and the same thing is applied for output pixel position (in case of other bit pixels, the pixel is equally divided into two for eg., for 16 bit pixel its divided into two parts with 8 bit pixel each). Now the first part containing the first four bits of the input pixel is swapped with the second part of the output pixel containing the last four bits and as well the other two parts are swapped and thus virtually re-creating a new pixel with new bit values.[4][5]

Example: consider input pixel position to be a_{22} and the corresponding output pixel position to be a_{53}. a_{22}= [10101101] and a_{53}= [01101001], where 'a' is the selected block. Initially a_{22} and a_{53} are as a_{22}= [{1010} {1101}] and a_{53}= [{0110} {1001}]. The intermediate steps are as follows: The pixels are broken down into individual values and are added separately into a new array and hence a swapping effect is created because the second part is actually swapped but the bits are added into the array individually in that form.

The snapshot of image and its encrypted image is shown below

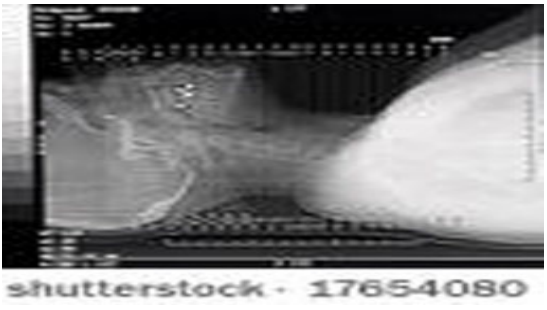


Fig. 4. Standard ct-scan image

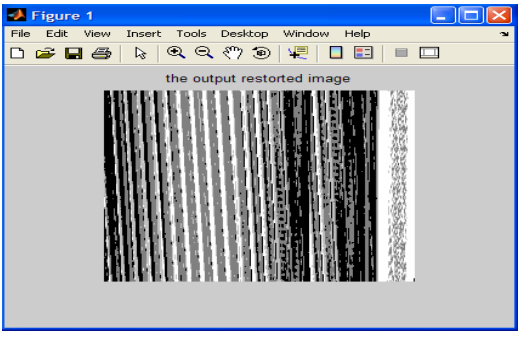


Fig. 5. Result of the encryption

Example: a_{22} is obtained as [1], [10], [101], [1010], [10101], [101010], [1010100], [10101001] where each one is an individual step.

Swapping the second part of a_{11} with the second part of a_{33} we get a_{22}= [{1010} {1001}] and a_{53}= [{0110} {1101}] which are virtually two new pixels with bit values a_{22}= [10101001] and a_{53}= [01101101] and thus by applying this process for the selected inputs as chosen by the user, the encryption is performed for those pixels in those particular positions and hence the finally ciphered or encrypted image is obtained. Symmetric key cryptography (the same secret key is used to encrypt and decrypt the image) is used which contains the information about the function used and the input's for which operation is performed. [6][7]

3.7 "Think" Captcha

Instead of fuzzy text we introduce a randomly generated real time image which will show some object or will portray some action. The user is expected to identify the action or the object. Questions are framed related to the picture by the human and as well the expected answers ie., the keywords it is commonly called as CAPTCHA. The work of the computer is to randomly throw these images with questions stored in the database to the user and compare the answers given by the users with the key words stored in the data base and conclude whether the user is a human or a robot based on the answer.

It is to be noted that the repetition of images will be avoided by removing the image its question and the related keywords from the database as soon as that particular entry is used. Though these types of systems already exists in graphical password system, [8] in graphical password system a set of options is given along with question to the user. In such cases there is a very high probability that the automated systems may select the correct answer based on permutations. But in this system, since the user is requested to type the answers, the work is not that easy for the automated systems whereas it is very easy for six sensed humans. Though the robots are facilitated with the help of artificial intelligence (A.I), the A.I of the present day systems is not that much developed to identify an action from a given real time image nor it can identify a real time image.[9] Our system is illustrated by the following example

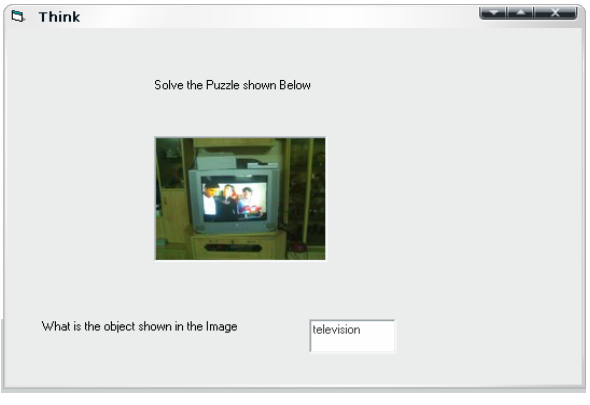


Fig. 6. Snapshot of THINK

The fig. 6 shows an image that is used in our system. It shows a person reading or studying a book. This is a real time image and the automated machines cannot guess the answer . Considering question, what is the object shown in the image? The key words (answer) are set as television, TV. If the user's answer contains any one of the above key words then it is identified that the user is a human and he is allowed to proceed further else it is identified as an automated system and the access is denied. [10][11]

4 Implementation and Experimental Results

The System developed uses the rule base to determine the disease name based on the symptoms entered by the user. This rule base consists of rules developed for a particular type of disease and determines the type of disease a patient has.

Based on the severity of the symptoms entered the system calculates the stage. Then it creates the ranges which represent the different stages of disease. The diagnosis logic based on these results (disease name and stage) and the rule base developed determines the appropriate diagnostic treatment for the patient.

Table 1. Shows a case of two patients diagnosed

Patient	Age	Symptom1	Symptom2	Symptom3	Symptom4	Diagnosis
Patient1	36	Seizures	Memory loss	Nausea	Loss of vision	Brain Tumor
Patient2	25	Yellow coloration	Itchy skin	Nausea	Vomiting	Hepatitis A/B

The rules populated in this rule base are developed for different type of diseases. These rules take into account the disease and the severity determined by the system to prescribe the treatment. The running time to diagnose a patient after the symptoms are entered is approximately around 1.28749 seconds. The implementation of the encryption algorithm was carried out using MATLAB. As the pool functions was mentioned a pool over 8 functions were used such as ex-or operation of two pixels selected followed by the exchange of last four bits using functions from the mentioned pool. The functions executed were similar to the Vernon cipher to obtain maximum efficiency in encryption.

Since Vernon cipher is the only efficient cipher or un-broken cipher, we derived functions similar to it so as to achieve the efficiency. This led to multiple sequential encryptions. The number of times a pixel being encrypted was counted and was maintained in fuzzy manner. Proper and organized decryption resulted in the original image. The running time of the respective encryption functions were 1.899002 and 1.899764 seconds. This shows the versatility of the system despite the variety functions, the average running time being 1.899829 seconds for 50 iterations. During Decryption, the similarity among the output images and the random pixels being not encrypted improved the system efficiency. Pentium core 2 duo processor with processing speed of 1.7 GHz was used with a 1 GB RAM. The running time of this image was around 22.79 seconds due to very high resolution. During Decryption, the similarity among the output images and the random pixels being not encrypted improved the system efficiency.

5 Conclusions and Future Works

Thus, the required system is developed which diagnoses users without a need to consult a doctor. Thus an expert system is developed. The system proposed is very simple and the implementation is very easy. This system is compatible with other techniques as the database can be used with case based reasoning and other mining methods. This capability of the system increases the scope and age of the system. This system is useful to the physician as well as the user for determining the type of disease and the stage of the disease.. The accuracy can be increased by implementing more analysis

techniques on the same database used in this system along with the current algorithm. The biggest advantage of this system is that it is secure which is not provided by other diagnosis systems also it provides THINK CAPTHCA and the encryption algorithm.

Computational study in secure medical diagnosis is still in its infancy stage and simpler but accurate and secure systems are the needed. The capability of this system for easy modification and continuous up gradation of the database is a plus point and increases the scope and life of the software. Here we have used Inference rules which can be replaced with fuzzy rules but the efficiency by doing so will not differ much but there is a chance that the running time of the system might increase. Thus we intend on developing a more efficient, sophisticated and secure system which can be used as current medical diagnosis systems and can be compatible with the mobile systems.

References

1. Discovering human understandable fuzzy diagnostic rules from medical data Giovanna Castellano, Anna Maria Fanelli and Corrado Mencar Department of Computer Science University of Bari
2. Saftoiu, A., Vilmann, P., Hassan, H., Gorunescu, F.: Analysis of endoscopic ultrasound elastography used for characterization and differentiation of benign and malignant lymph nodes. Ultraschall in der Medizin (European Journal of Ultrasound) 27(6), 535–542 (2006)
3. Tsakonas, A., Dounias, G., Jantzen, J., Axer, H.: Evolving rule based systems in two medical domains using genetic programming. Artificial Intelligence of Medicine- Elsevier Journal 32, 195–216 (2004)
4. Alsultanny, Y.A.: Image encryption by Cipher Feedback Mode, Computer Engineering Department Applied Science University, Amman 11931, Jordan. International Journal of Innovative Computing, Information and Control 3(3) (June 2007)
5. Zheng, W.M., Lu, Z.M., Burkhardt, H.: Color image retrieval schemes using index histograms based on various spatial-domain vector quantizers. International Journal of Innovative, Computing, Information & Control 2(6), 1317–1327 (2006)
6. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice-Hall, Inc., Englewood Cliffs (2002)
7. Yahya, A.A., Abdalla, A.M.: A Shuffle Image-Encryption Algorithm Department of Computer Science, Al-Zaytoonah University of Jordan, Amman 11733, Jordan
8. Jablon, D.: Strong password-only authenticated key exchange. Computer Communication Review 265, 5 (1996)
9. Srikanth, V., Vishwanathan, C., Asati, U., Sriman Narayana Iyengar, N.C.: THINK (Testifying Human Based On Intelligence and Knowledge). In: Proceedings of ICAC 2009, Mumbai, Maharashtra, India, January 23-24 (2009) Copyright 2009 ACM 978-1-60558-351-8
10. Bellovin, S.M., Merritt, M.: Encrypted key exchange: Password-based protocols secure against dictionary attacks. In: Proceedings of the 1992 IEEE Computer Society Conference on Research in Security and Privacy, pp. 72–84 (1992)
11. Gong, L., Lomas, M.A., Needham, R., Saltzer, J.: Protecting poorly chosen secrets from guessing attacks. IEEE Journal on Selected Areas in Communications 115, 648–656 (1993)

Factors Affecting Intention to Use in Social Networking Sites: An Empirical Study on Thai Society

Rath Jairak[1], Napath Sahakhunchai[2], Kallaya Jairak[1], and Prasong Praneetpolgrang[2]

[1] Doctor of Philosophy Program in Information Technology
[2] Master of Science Program in Information Technology
Information Science Institute, Sripatum University, Bangkok, Thailand 10900
{rathjairak,napathpk,ajkallaya}@gmail.com, prasong.pr@spu.ac.th

Abstract. This research aims to explore the factors that affect the intention to use in Social Networking Sites (SNS). We apply the theory of Technology Acceptance Model (TAM), intrinsic motivation, and trust properties to develop the theoretical framework for SNS users' intention. The results show that the important factors influencing SNS users' intention for general purpose and collaborative learning are task-oriented, pleasure-oriented, and familiarity-based trust. In marketing usage, dispositional trust and pleasure-oriented are two main factors that reflect intention to use in SNS.

Keywords: Social networking sites, Online social activity, Intention to use, SNS adoption.

1 Introduction

Social Networking Sites (SNS) such as Myspace, Facebook, and Hi5 are online communities that have attracted millions of users, many of them have admitted SNS activity as a part of everyday life. The major tasks of SNS aim to provide their users with social activity such as invitation for new friends, connection to old friends, distribution of private/public contents, and discussion of social issues. In SNS, users can join, create, and share information within and among groups. The components in SNS should serve for both task-oriented and human-relationship oriented service [1].

Most of online service providers normally aim to provide useful information with the high level of system usability for users. Even in the service that available for human-relationship in SNS, task-oriented properties are also the basic configuration for this system to ensure that users can reach the high level of performance in task when they intend to participate with social activity in SNS. Technology Acceptance Model (TAM) was first introduced in [2]. It has been widely used to describe the traditional task-oriented information system adoption. TAM can also be integrated with trust concepts to predict the human-trust-oriented information system adoption (i.e. e-Commerce) [3, 4]. But it is hard to identify the rigorous model to give a holistic explanation for new information system concept like social networking community.

This study is one of the few attempts to investigate the factors that affect intention to use SNS in Thai society. We conduct a series of statistical analysis to explore

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 43–52, 2010.

which individual beliefs that affect SNS users' intention in three difference contexts: usage for general purpose, usage for collaborative learning, and usage for marketing communication.

The rest of this paper is organized as follows: in section 2, we review the research background and research model, our hypotheses also suggest in this section. In section 3, we describe about the research method. The results and discussion are summarized in section 4 and 5. Finally, in section 6, we conclude our limitation and future research.

2 Theoretical Background and Research Model

In this section, we briefly discuss some related theoretical background and research model that we propose to conduct our hypotheses.

2.1 Task-Oriented Properties

TAM has been widely used to examine user adoption and continuance usage intention in wide range of information system such as e-Mail adoption [2], Internet adoption [5], continuance of purchasing from e-Commerce [3, 4], e-Learning adoption [6], continuance usage of e-Government website [7], and social network service adoption [1]. Based on the rigour of TAM theory, we would expect that perceived usefulness and perceived each-of-use, the two major constructs in TAM, have the power to predict SNS adoption. Perceived usefulness is defined as "the degree to which a person believes that using a particular system would enhance his or her job performance" [2]. The second construct, perceived each-of-use is defined as "the degree of to which a person believes that using a particular system would be free of effort" [2].

2.2 Pleasure-Oriented Property

Many studies in the past have shown that intention to use information system involve both extrinsic (i.e. perceived usefulness) and intrinsic (i.e. perceived enjoyment) motivation [4, 5, 6]. Extrinsic motivation is defined as "the performance of an activity because it is perceived to be instrumental in achieving valued outcomes that are distinct from the activity itself" [5]. Intrinsic motivation is defined as "the performance of an activity for no apparent reinforcement other than the process of performing the activity per se." [5]. Extrinsic motivation depends on the instrumental value in the activity [8]. For example, People work hard because they believe that they will get more benefit when doing this. In contrast, intrinsic motivation is reflected from inside rather than from any external or outside rewards. Perceived enjoyment is intrinsic motivation and we can postulate that enjoyment will play a significant role in SNS users' intention.

2.3 Human-Trust-Oriented Properties

Trust is a prerequisite for human-computer interaction, especially within the information system that trustor feel vulnerable to deal with trustee. Trust is studied in the field

of human-trust-oriented information system such as e-Commerce [3, 4, 9] and knowledge sharing in virtual community [10, 11, 12]. In this study, we aim to uncover all trust properties and find out which properties have a significant effect on SNS users' intention. We have applied the Kim's concept of trust from e-Commerce entities in [9]. Based on this concept, trust consist of 4 properties: Cognitive-based (privacy and security concerns), Affect-based (indirect interactions with trustee such as reputation), Experience-based (familiarity with the system usage) and Personality-based (personal belief in willingness to depend on others in community).

2.4 Research Model

This study employs the theory of TAM, intrinsic motivation, and trust properties to develop the theoretical framework for SNS users' intention. The research model that we intend to explore is shown in Fig. 1.

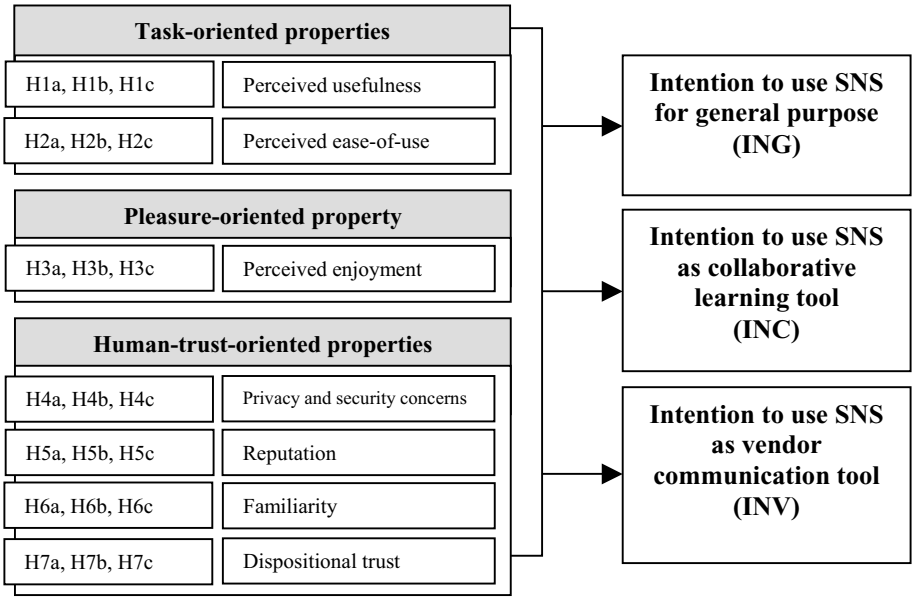


Fig. 1. The research model of SNS users' intention

2.5 Research Hypothesis

A set of hypotheses were developed based on theoretical background and the aim of this research, we classified 21 hypotheses into three parts.

Part 1, Task-oriented properties: The relationship of perceived usefulness and perceived ease-of-use toward SNS users' intention.

H1a. Perceived usefulness will have a positive effect on ING.
H1b. Perceived usefulness will have a positive effect on INC.

H1c. Perceived usefulness will have a positive effect on INV.
H2a. Perceived ease-of-use will have a positive effect on ING.
H2b. Perceived ease-of-use will have a positive effect on INC.
H2c. Perceived ease-of-use will have a positive effect on INV.

Part 2, Pleasure-oriented property: The relationship of perceived enjoyment toward SNS users' intention.

H3a. Perceived enjoyment will have a positve effect on ING.
H3b. Perceived enjoyment will have a positive effect on INC.
H3c. Perceived enjoyment will have a positive effect on INV.

Part 3, Human-trust-oriented properties: The relationship of security and privacy concerns, reputation, familiarity, and dispositional trust toward SNS users' intention.

H4a. Privacy and security concerns will have a positive effect on ING.
H4b. Privacy and security concerns will have a positive effect on INC.
H4c. Privacy and security concerns will have a positive effect on INV.
H5a. Reputation will have a positive effect on ING.
H5b. Reputation will have a positive effect on INC.
H5c. Reputation will have a positive effect on INV.
H6a. Familiarity will have a positive effect on ING.
H6b. Familiarity will have a positive effect on INC.
H6c. Familiarity will have a positive effect on INV.
H7a. Dispositional trust will have a positive effect on ING.
H7b. Dispositional trust will have a positive effect on INC.
H7c. Dispositional trust will have a positive effect on INV.

3 Research Methods

This paper aims to explore the factors that affect SNS users' intention. Our questionnaire consists of three main parts. The first part is about demographic profile while the second part is related to usage behavior in SNS and the third part focus on the factors that affect usage intention. We used non-probability sampling to collect questionnaires from 300 persons who had used SNS. After the survey was completed, a series of statististical analysis was conducted to test the hypotheses. In this section, we describe the necessary details in our research methods.

3.1 Measurement Development

Measurement items were developed from the source of literature, as shown in Table1. All 33 items were carefully selected, adapted, and translated to Thai language via many rounds of discussion in our research team. We developed two new constructs (Intention to use SNS as collaborative learning tool and Intention to use SNS as vendor communication tool) to provide a better understanding of behavioral intention for

Table 1. Development of SNS questionnaire

Context	Source of literature
Task-oriented	
Perceived usefulness	Kwon and Wen (2010)
Perceived ease-of-use	Kwon and Wen (2010)
Pleasure-oriented	
Perceived enjoyment	Chiu, Chang, and Cheng (2009)
Human-trust-oriented	
Privacy and security concerns	Hsu et al. (2007)
Reputation	Kim, Ferrin, and Rao (2008)
Familiarity	Kim, Ferrin, and Rao (2008)
Dispositional trust	Kim, Ferrin, and Rao (2008)
Intention to use	
Intention to use SNS for general purpose	Kwon and Wen (2010)
Intention to use SNS as collaborative learning tool	New construct
Intention to use SNS as vendor communication tool	New construct

SNS. All items were measured using a positive word questions on a 5-point Likert scale from 1-strongly disagree to 5-strongly agree.

3.2 Data Collection

The data were collected from 300 respondents and there were 229 usable responses, yieding a response rate of 76.3%. 49% of respondents were males and 51% were females. Based on age and educational status, the majority (>60%) of the respondents were in the educational age. Around 56% spend 1-5 hours for SNS daily usage. Table 2 shows the demographic details of the respondents.

Table 2. Demographic details (the number of subjects = 229)

Variable	Value	Frequency	Percentage
Gender	Male	113	49%
	Femal	116	51%
Age (years)	< 20	101	44%
	20-25	39	17%
	26-30	46	20%
	> 30	43	19%
Occupation	Public sector	31	13%
	Private sector	59	26%
	Student	139	61%
Educational status	Under bachelor degree	94	41%
	Bachelor degree	89	39%
	Graduate school	46	20%
Daily SNS usage (hours)	< 1	74	32%
	1-5	128	56%
	> 5	27	12%

4 Data Analysis and Results

In this study, we employed a factor analysis to confirm convergent and discriminant validity in our instrument. And performed a regression analysis to test for hypotheses and find causal relationships between dependent and independent variables in the research model.

Table 3. Summary of SNS measurement scales

Construct	Mean	SD	Factor loading	Cronbach's alpha
Task-oriented and Familiarity (TOFAM)				***0.889***
TOFAM1	3.89	0.83	0.70	
TOFAM2	3.85	0.90	0.72	
TOFAM3	3.91	0.93	0.69	
TOFAM4	3.92	0.80	0.58	
TOFAM5	3.77	0.82	0.57	
TOFAM6	3.89	0.83	0.61	
TOFAM7	3.88	0.79	0.61	
TOFAM8	4.00	0.86	0.74	
TOFAM9	3.98	0.77	0.61	
Enjoyment (EN)				**-**
EN	3.93	0.80	0.58	
Privacy and Security Concerns (PSC)				***0.753***
PSC1	3.34	0.94	0.80	
PSC2	3.32	1.00	0.74	
PSC3	3.66	0.98	0.74	
Reputation (REP)				***0.829***
REP1	4.47	0.77	0.79	
REP2	4.28	0.86	0.71	
REP3	4.20	0.85	0.70	
Dispositional trust (DT)				***0.823***
DT1	3.08	0.91	0.75	
DT2	3.01	0.97	0.71	
Intention to use SNS for general purpose (ING)				***0.870***
ING1	3.57	1.00	0.78	
ING2	3.01	1.00	0.71	
ING3	3.61	0.96	0.76	
ING4	3.59	0.97	0.63	
Intention to use SNS as collaborative learning tool (INC)				***0.795***
INC1	3.75	0.86	0.77	
INC2	3.66	0.89	0.75	
INC3	3.65	0.98	0.57	
Intention to use SNS as vendor communication tool (INV)				***0.878***
INV1	2.70	1.20	0.78	
INV2	3.06	1.10	0.86	
INV3	2.91	1.17	0.88	

Table 3 shows the factor analysis and reliability results. The factor loading was conducted with varimax rotation. Twenty-eight items remained after five were dropped due to low factor loadings (below 0.5) [13, 14]. Our initial instrument could

not extract the components in perceived usefulness, perceived ease-of-use, and familiarity based trust, thus all components were grouped into task-oriented and familiarity (TOFAM). Two items in perceived enjoyment were suppressed, there was only one item that represent the perceived enjoyment construct. All factors had Cronbach's alpha above 0.7, it was indicated that internal consistency can be obtained [15, 16].

Table 4. Regression Analysis Results

Independent variables	Dependent variables								
	Intention to use SNS for general purpose (ING) ($R^2 = 0.339$)			Intention to use SNS as collaborative learning tool (INC) ($R^2 = 0.304$)			Intention to use SNS as vendor communication tool (INV) ($R^2 = 0.146$)		
	Beta	t	Sig.	Beta	t	Sig.	Beta	t	Sig.
TOFAM	.427	5.882	.000**	.365	4.896	.000**	.096	1.162	.247
EN	.245	3.739	.000**	.299	4.449	.000**	.196	2.629	.009**
PSC	.055	.886	.377	.043	.665	.506	-.035	-.493	.622
REP	-.860	-1.241	.216	-.153	-2.161	.032*	-.107	-1.363	.174
DT	.053	.838	.403	.085	1.307	.193	.267	3.717	.000**

*Correlation are significant at *p <.05, **p <.01*

We performed a regression analysis to test for hypotheses, the research model was analyzed with three separate dependent variables: usage for general purpose (ING), usage for collaborative learning (INC), usage for marketing communication (INV). Table 4 presents the results of regression analysis.

In the case of ING, TOFAM and EN were found to be significant to ING, and explained 34 percent of ING variance. However, PSC, REP, and DT were found to be insignificant. Thus, H1a, H2a, H3a, and H6a were supported. The beta value of TOFAM (Beta = 0.427) was greater than EN (Beta = 0.245), TOFAM had more impact on users' intention than EN. These results indicated that the high level of usefulness, ease-of-use, familiarity and enjoyment can enhance SNS users' intention for general usage.

For the case of INC, TOFAM and EN were found to be significant to INC, and explained 30 percent of INC variance. However, PSC and DT were found to be insignificant. Surprisingly, REP was found to have a significant negative effect on INC. Therefore, H1b, H2b, H3b, and H6b were supported. The beta value of TOFAM (Beta = 0.365) was greater than EN (Beta = 0.299), TOFAM had more impact on users' intention than EN. These results indicated that the high level of usefulness, ease-of-use, familiarity, and enjoyment can enhance SNS users' intention for collaborative learning tool. In addition, based on a negative effect of reputation, we can conclude that even well known SNS may not provide a complete environment for callaborative learning activity.

For the last case (INV), EN and DT were found to be significant to INV, and explained 15% of INV variance. However, TOFAM, PSC, and REP were found to be insignificant. Thus, H3c and H7c were supported. The beta value of DT (Beta = 0.267) was greater than EN (Beta = 0.196), DT had more impact on users' intention

than EN. These results indicated that the high level of dispositional trust and enjoyment can enhance SNS users' intention for vendor communication tool.

Based on the survey responses, there were 86 respondents who have used Facebook regularly, 78 have used Hi5 regularly, 12 have used twitter regularly, and 5 have used Myspace regularly. For the proper interpretation in the number of comparison, we identified only Facebook and Hi5 users. The comparison results are plotted in Fig. 2. The mean value for each factors was interpreted as percentage to allow for clearer comparison. The Facebook users rated for overall factors with higher score than Hi5 users, except only the privacy and security concerns that Hi5 retrieve higher score than Facebook. Our study also found that Facebook and Hi5 are two of the most popular SNS among respondents. Workers engaged to use Facebook more than Hi5. For students, they used Hi5 more than Facebook. Even Hi5 was the most popular among students, it also was the most cancelled among students and workers.

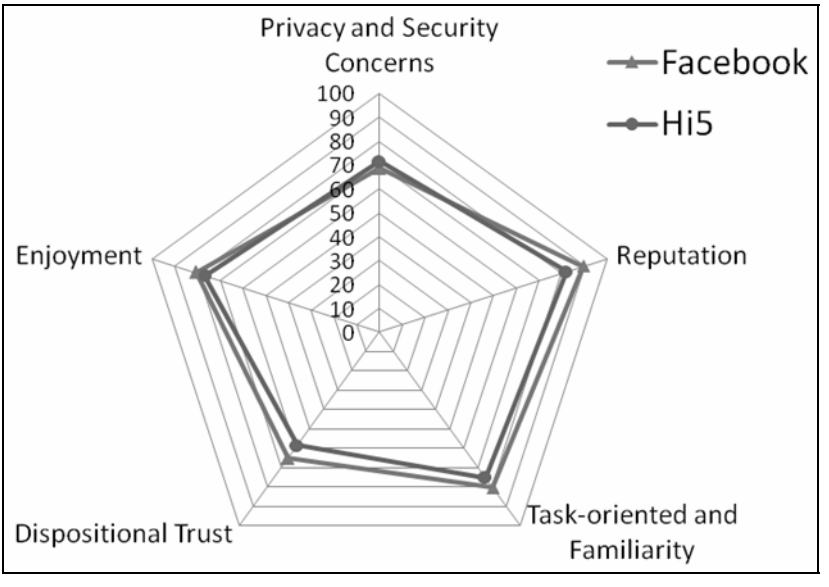


Fig. 2. Radar chart of affecting factors for two SNS

5 Discussion

In this paper, we focus on the factors that affect SNS users' intention in three different contexts: usage for general purpose, usage for collaborative learning, and usage for marketing communication. The results indicate that the main factors that influence SNS users' intention for general purpose and collaborative learning are task-oriented, pleasure-oriented, and familiarity-based trust. For marketing usage, dispositional trust and pleasure-oriented are two main factors that reflect intention to use in SNS. Furturemore, we have found that security and privacy concerns are not relevant with SNS users' intention in all different contexts. Users still use in SNS that provide the

well-fitting benefits for online social activity, even when security and privacy safeguards are weak [17].

For general usage in SNS, task-oriented properties are the basic configuration to serve human interaction in online social activity such as invitation for new friends, connection to old friends, distribution of private/public contents, and discussion of social issues. Change in use of SNS provider is caused by the benefits in online social activity that participants perceive from the system. A recent study also suggested that SNS users' intention is more strongly related to social factors than task relevant [18]. Thus, the key success factors for SNS that provide general service are associated with the quality of system usablitity and attraction in online social activity.

For collaborative learning in SNS, the social software that serve for learning acitivity should provide the collaborative features such as distributing useful content, editing content with co-authoring, and tagging the trusted material for sharing within and among groups [19]. Thus, the well know SNS in this day may not provide a complete environment for callaborative learning activity, this is the opportunity to investigate more which SNS that provide a suitable environment for learning activity or the academic professionalism itself should contribute a new social software for collaborative learning tool.

For marketing usage, trusting other people and fun activity that provide by vendor community are the keys to leading customers intention to interact with online vendors through SNS. Transaction with familiar partners also increase satisfaction among customers [20]. The marketing activity in SNS should apply with viral marketing to increase brand awareness and integrate with high entertainment value of activity to create customers value [21].

6 Limitation and Future Research

There are some significant limitations to this study. Firstly, the instrument could not extract some significant components (perceived usefulness, perceived ease-of-use and familiarity based trust) that may be caused by the translation process or confusion among the participants with the related concepts in usefulness, ease-of-use, and familiarity. Secondly, the social factors (i.e. perceived encouragement in [1], collectiveness in [18]) were not added to investigate in our model. The future researchs should focus more on social factors that possible related to SNS users' intention and also investigate more deeply with the rigorous surveys in a variety of services in SNS. In addition, game-based motivation in SNS is also an interesting issue to find more expression, especially in the context of marketing and collaborative learning services in social networking sites.

References

1. Kwon, O., Wen, Y.: An empirical study of the factors affecting social network service use. Computers in Human Behavior 26, 254–263 (2010)
2. Davis, F.D.: Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly 13, 319–340 (1989)

3. Gefen, D., Karahanna, E., Straub, D.W.: Trust and TAM in Online Shopping: An Integrated Model. MIS Quarterly 27, 51–90 (2003)
4. Chiu, C.M., Chang, C.C., Cheng, H.L., Fang, Y.H.: Determinants of customer repurchase intention in online shopping. Online Information Review 33, 761–784 (2009)
5. Teo, T.S.H., Lim, V.K.G., Lai, R.Y.C.: Intrinsic and extrinsic motivation in Internet usage. Omega 27, 25–37 (1999)
6. Lee, M.K.O., Cheung, C.M.K., Chen, Z.: Acceptance of Internet-based learning medium: the role of extrinsic and intrinsic motivation. Information & Management 42, 1095–1104 (2005)
7. Wangpipatwong, S., Chutimaskul, W., Papasratorn, B.: Understanding Citizen's Continuance Intention to Use e-Government Website: a Composite View of Technology Acceptance Model and Computer Self-Efficacy. Electronic Journal of e-Government 6, 55–64 (2008)
8. Ryan, R.M., Deci, E.L.: Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. Contemporary educational psychology 25, 54–67 (2000)
9. Kim, D.J., Ferrin, D.L., Rao, H.R.: A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. Decision Support Systems 44, 544–564 (2008)
10. Chiu, C.M., Hsu, M.H., Wang, E.T.G.: Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. Decision Support Systems 42, 1872–1888 (2006)
11. Hsu, M.H., Ju, T.L., Yen, C.H., Chang, C.M.: Knowledge sharing behavior in virtual communities: The relationship between trust, self-efficacy, and outcome expectations. International Journal of Human-Computer Studies 65, 153–169 (2007)
12. Zhang, Y., Fang, Y., Wei, K.K., Chen, H.: Exploring the role of psychological safety in promoting the intention to continue sharing knowledge in virtual communities. International Journal of Information Management (2010) doi: 10.1016/ijinfomgt.2010.02.003
13. Bontis, N.: Intellectual capital: an exploratory study that develops measures and models. Management decision 36, 63–76 (1998)
14. Watjatrakul, B., Barikdar, L.A.: Attitudes toward Using Communication Technologies in Education: A Comparative Study of Email and SMS. In: Third International Conference, Advances in Information Technology, pp. 191–201 (2009)
15. Gliem, J.A., Gliem, R.R.: Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. In: The Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, Ohio (2003)
16. Hair, J.F., Money, A.H., Samouel, P., Page, M.: Research methods for business, pp. 242–245. John Wiley and Sons, Chichester (2007)
17. Dwyer, C., Hiltz, S.R., Passerini, K.: Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. In: Proceedings of AMCIS (2007)
18. Hunprasert, P., Vorakulpipat, C., Siwamogtham, S.: A social impact of the use of social networking site: A case study of Hi5. NECTEC Technical Journal 21, 230–238 (2009)
19. Liccardi, I., Ounnas, A., Pau, R., Massey, E., Kinnunen, P., Lewthwaite, S., et al.: The role of social networks in students' learning experiences. ACM SIGCSE Bulletin. 39, 224–237 (2007)
20. Swamynathan, G., Wilson, C., Boe, B., Almeroth, K., Zhao, B.Y.: Do social networks improve e-commerce?: a study on social marketplaces. In: Proceedings of the first Workshop on Online Social Networks, pp. 1–6 (2008)
21. Preibusch, S., Hoser, B., Gürses, S., Berendt, B.: Ubiquitous social networks–opportunities and challenges for privacy-aware user modelling. In: Proceedings of the Data Mining for User Modelling Workshop, Corfu (2007)

A General Bayesian Network Approach to Analyzing Online Game Item Values and Its Influence on Consumer Satisfaction and Purchase Intention

Kun Chang Lee[1] and Bong-Won Park[2]

[1] Professor of MIS and WCU Professor of Creativity Science
SKK Business School and Department of Interaction Science
Sungkyunkwan University
Seoul 110-745, Republic of Korea
kunchanglee@gmail.com, leekc@skku.edu
[2] Department of Interaction Science, Sungkyunkwan University
Seoul 110-745, Republic of Korea
combio00@naver.com

Abstract. Many online game users purchase game items with which to play free-to-play games. Because of a lack of research into which there is no specified framework for categorizing the values of game items, this study proposes four types of online game item values based on an analysis of literature regarding online game characteristics. It then proposes to investigate how online game users perceive satisfaction and purchase intention from the proposed four types of online game item values. Though regression analysis has been used frequently to answer this kind of research question, we propose a new approach, a General Bayesian Network (GBN), which can be performed in an understandable way without sacrificing predictive accuracy. Conventional techniques, such as regression analysis, do not provide significant explanation for this kind of problem because they are fixed to a linear structure and are limited in explaining why customers are likely to purchase game items and if they are satisfied with their purchases. In contrast, the proposed GBN provides a flexible underlying structure based on questionnaire survey data and offers robust decision support on this kind of research question by identifying its causal relationships. To illustrate the validity of GBN in solving the research question in this study, 327 valid questionnaires were analyzed using GBN with what-if and goal-seeking approaches. The experimental results were promising and meaningful in comparison with regression analysis results.

Keywords: Customer retention, Online games, Game item values, Bayesian network, General Bayesian Network, Causal relationships.

1 Introduction

In recent years, the online game market has grown exponentially along with an increase in Internet usage [1, 2, 3]. Online game users customarily pay a monthly fee to play subscription-based games. This type of business model was popular in the early

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 53–62, 2010.

phase of online games. However, free-to-play games are currently more popular than are subscription-based games [4]. The free-to-play business model provides free access to basic game features, but online game users have the option to purchase game items for several gameplay purposes; these items either enhance their power in the game's context or improve their game characters' visual appearance. Game users may register at no cost for free-to-play games, but if they want to decorate their characters and/or empower them for a special purpose in the game context, they are required to purchase the relevant game items [4, 5]. Despite a recent surge in free-to-play game popularity, no significant study exists in the literature to clearly answer, with sufficiently convincing analysis, how online game item values and related factors influence game users' purchase intentions. This study seeks to address that research question.

Traditionally, the regression method has been applied to answer this kind of research question. However, regression poses serious problems, such as its rigid functional form. The linear form of the regression equation limits the interpretability of regression results. To overcome this problem, this study proposes a new approach that uses the General Bayesian Network (GBN) [6, 7]. The GBN is unique in the sense that its structure mirrors the causal relationships between a class node (target variable) and all other nodes. When a dataset is available in the target problem domain, the GBN structure can be induced [6] so that the GBN enables decision makers to simulate what-if and goal-seeking analyses so as to investigate a hidden cause underlying the target problem. Therefore, this study adopts the GBN to induce a causal relationship between four game item values, character identification, satisfaction and a customer's purchase intention.

2 Previous Studies

2.1 Theory of Consumption Values and Game Item Values

Game item values are perceived differently by different game users, depending on their personal tastes. In other words, game items are consumer goods that are available in virtual environments [8]. The question of how game users value game items should be approached from an economic perspective. In this sense, this study investigates game item values using the theory of consumption values [9]. According to the theory of consumption values, consumers perceive the values of the target goods with respect to five dimensions: functional value, social value, emotional value, conditional value, and epistemic value. This theory is widely used to analyze consumer behavior, such as with airline frequent flyer programs [10], clothing [11], organic food [12], sponsorship in sports marketing [13], and tourism [14]. Furthermore, it does not apply only to traditional products and services, but also to new IT product and services, such as smart phones [15], mobile Internet service [16].

Meanwhile, since game items are belonging to digital goods, we have to consider other applications where the theory of consumption values has been applied to other kinds of digital goods. Turel et al. [17] explained how users adopt ringtone in mobile phone using the theory of consumption values. They found that there exist four kinds of ringtone values such as visual/musical appeal value, social value, playfulness value, and value for money, and that the three values except social value affect the

overall value of the ringtone in mobile phone. Such overall value of ringtone in mobile phone is known to affect users' intention to use and positive word-of-mouth.

Game item values belong to virtual goods that are available only in game environments. Therefore, to understand the purchase of online game items in free-to-play games, it is necessary to review related research studies. From this review, we hoped to determine why game users purchase the game items. To this end, Lin and Sun [4] categorized game items into two kinds of game items: functional props and decorative props. Second, game items are classified into three types -- vanity, functional, and social items -- according to a paper from Live Gamer [18]. Third, Guo and Barnes [8] determined the motivations for purchasing virtual items: perceived playfulness, character competency, and the requirements of the quest system. Fourth, Lehdonvirta [19] found that virtual goods have three kinds of attributes: functional, emotional, and social. Functional attributes include performance and functionality. Hedonic and social attributes pertain to visual appearance and sounds, background fiction, provenance, customizability, cultural references, branding, and rarity.

To derive the appropriate dimensions of the game item values based on the theory of consumption values and the characteristics of game items in the game literature, we suggest the following four values of game items. The first value is "character competency value," from which game users obtain a performance advantage by using the game item (e.g., "10% faster"). The second value is "enjoyment value," from which game users derive fun and escape from the stress of the real world. The third value is "visual authority value," from which game users achieve an aesthetic effect and reap social gains from other game users. The fourth value is "monetary value," the user-perceived cost-effectiveness of a game item.

2.2 Other Factors Affecting Purchase Intention

This study seeks to investigate how the four dimensions of game item values and the two factors of character identification and satisfaction influence game users' purchase intentions. Since the four dimensions of the game item values were explained in the previous section, it is necessary to describe a related literature review for the two factors of character identification and satisfaction.

Character identification. In media research, Hefner et al. [20] defined identification with a media character as "a temporary alteration of the media users' self-perception by including the perceived properties of the target media character."

When a game user plays a specific online game for the first time, a game character must be created. When a game player purchases online game items, these items are applied to this virtual character. In particular, previous research [20, 21] has shown that game users identify and form quite a close relationship with their game characters.

Satisfaction. Customer satisfaction is a continual issue for marketing research and is an important factor that affects purchase intention [22]. In general, satisfied customers have re-purchase (visit) intention. Similarly, when game users are satisfied with an online game, they continue playing it and seek to purchase related game items, which seem to improve the values users seek to obtain.

2.3 Bayesian Network

A Bayesian network (BN) is a graphical model of a probability distribution assigning a probability to an event of interest [23, 24]. Basically, a BN is a directed acyclic graph in which the nodes correspond to domain variables $x_1, \ldots, x_n$ and the arcs between nodes represent direct dependencies between variables. Consequently, we should be aware that the BN structure extracted from a specific problem domain shows an underlying causal relationship among variables in the analysis, and that such a causal relationship represented in the BN structure provides important information with which decision makers can make an informed decision. A BN can be induced from a dataset collected from a specific problem domain. In this case, the BN must specify a class node and other explanatory nodes [25]. Induction of the BN structure from the dataset is based on the assumptions that all variables are discrete and observed and that each dataset occurs independently and has no missing values. However, the BN can also be used as a classifier when users want to determine whether the exact probability of an event is above (or below) a certain threshold [26]. When the BN is used as a classifier, a class node should be designated in advance. The class node then becomes a target node with which other nodes are interlinked depending on the structure.

To improve classification performance, Cheng and Greiner [6] suggested two BN structures called a Bayesian Network Augmented NBN (BAN) and a General Bayesian Network (GBN). A BAN allows all other nodes to be direct children of the class node, but a complete BN is constructed between the child nodes. Meanwhile, a GBN is a full-fledged BN in which causal relationships between the class node and all other nodes are flexibly formulated using an efficient network construction technique based on conditional independence tests [27].

3 Research Methodology and Experiment

3.1 Questionnaire Survey and Sample Statistics

To analyze the relationships among factors affecting purchase intention, survey items were adapted from previous studies. For example, in case of enjoyment value, three survey items were made from Turel et al. [17] and Guo & Barnes [8]. To determine character identification, four survey items were developed from Cohen [28] and Hefner et al. [20]. After constructing the questionnaire, we randomly selected online game users in South Korea. The number of initial respondents was 384, and after we assessed the response quality, 327 responses were selected as suitable for this study, of which 250 were from males and 77 were from females. Respondents consisted of 131 middle school, 117 high school, and 79 college students.

3.2 Reliability and Confirmatory Factor Analysis

Seven constructs were used in the research question, namely, enjoyment value, character competency value, visual authority value, monetary value, character identification, satisfaction, and purchase intention. In Table 1, Cronbach's alpha values for the

seven constructs were all greater than 0.7, indicating that these items were reliable. In addition, principal component analysis with the varimax rotation option was used to test the validity of each item. The first factor, satisfaction, explained 32.3% of the total variance; the second factor, character identification, 12.3%. The total variance explained by the seven factors was 75.4%. On the basis of these results, we concluded that the questionnaire items were statistically valid.

Table 1. Reliability and Factor analysis

Variables	Cronbach's alpha	SA	CI	VV	PI	CC	EN	MO
satis4	0.91	**0.89**	0.05	0.09	0.09	0.05	0.12	0.02
satis3		**0.88**	0.06	0.09	0.07	0.11	0.09	0.04
satis2		**0.85**	0.03	0.10	0.07	0.15	0.10	0.02
satis1		**0.84**	0.06	-0.07	0.09	0.08	0.20	0.11
charaiden4	0.87	0.02	**0.87**	0.10	0.12	0.03	0.00	0.12
charaiden2		0.02	**0.84**	0.12	0.17	0.03	0.06	0.13
charaiden1		0.04	**0.80**	0.18	0.07	0.15	0.02	0.12
charaiden3		0.11	**0.76**	0.05	0.17	0.05	0.11	0.06
visual3	0.84	0.02	0.10	**0.80**	0.22	0.01	0.19	0.15
visual2		0.03	0.10	**0.79**	0.16	0.06	0.09	0.05
visual4		0.08	0.14	**0.76**	0.15	0.07	0.14	0.27
visual1		0.10	0.13	**0.70**	0.09	0.22	0.10	0.09
intention3	0.93	0.11	0.21	0.21	**0.88**	0.12	0.07	0.14
intention1		0.14	0.17	0.16	**0.85**	0.15	0.11	0.14
intention2		0.10	0.23	0.26	**0.83**	0.08	0.13	0.13
competency2	0.78	0.13	0.06	0.04	0.17	**0.84**	0.12	0.07
competency1		0.13	0.03	0.09	0.09	**0.81**	0.23	0.09
competency3		0.10	0.16	0.20	0.05	**0.66**	0.15	0.21
enjoyment1	0.84	0.20	0.06	0.06	0.18	0.11	**0.83**	0.06
enjoyment2		0.19	0.04	0.27	0.09	0.29	**0.78**	0.06
enjoyment3		0.20	0.12	0.32	0.01	0.26	**0.70**	0.20
money2	0.76	0.10	0.13	0.19	0.12	0.21	0.10	**0.82**
money3		0.01	0.17	0.10	0.11	0.09	-0.02	**0.82**
money1		0.08	0.14	0.27	0.18	0.08	0.31	**0.60**
Eigenvalues		7.75	2.96	2.10	1.61	1.41	1.25	1.01
Variance explained (%)		32.3	12.3	8.7	6.7	5.9	5.2	4.2
Total variance explained (%)		32.3	44.6	53.4	60.1	65.9	71.2	75.4

Note: SA: Satisfaction, CI: Character Identification, VV: Visual Authority Value, PI: Purchase Intention, CC: Character Competency Value, EN: Enjoyment Value, MO: Monetary Value

3.3 Results

To apply the GBN mechanism to the questionnaire data, we calculated the average values for each factor and then transformed the Likert scale for each factor into either Low (1 to 3), Medium (3 to 5), or High (5 to 7). Causal relationships among the seven variables were depicted by the GBN result, as shown in Figure 1.

First, the five variables of enjoyment value, visual authority value, character competency value, character identification, and satisfaction were found to have a direct relationship to purchase intention. Second, visual authority value is related to enjoyment value and monetary value. Likewise, enjoyment value is associated with character competency value and satisfaction, and monetary value is related to visual authority value and character competency value. The prediction accuracy of this model was 60.6 %. Consequently, casual relationships are able to be easily identified using GBN. In addition, the most important node in relation to the 'purchase intention' node is character identification (relative significance: 1.00), followed by satisfaction (0.83), visual authority value (0.81), and character competency value (0.61).

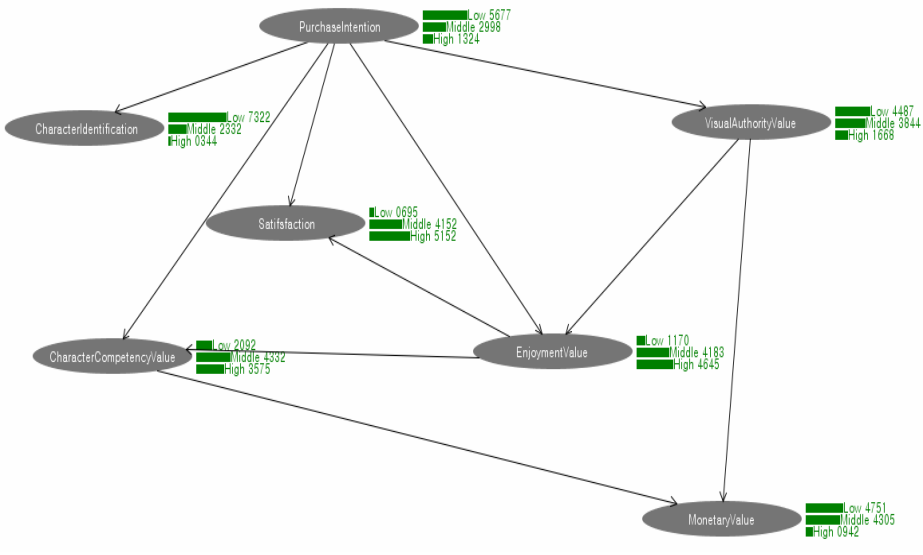


Fig. 1. GBN with purchase intention as a target node

In order to know the what-if and goal-seeking support capability of the GBN, let us consider the following two scenarios, in which sensitivity analysis is performed by taking advantage of causal relationships suggested by Figure 1.

Scenario 1 (What-If analysis): Among the four values, if the visual authority value is high (i.e., it has a value between 5 and 7) and no other variables are changed, how do the purchase intention and other variables change?

As shown in Figure 2, the visual authority value is related with purchase intention, enjoyment value, and monetary value. Originally, the prior probability of low purchase intention has the largest probability. However, when the visual authority value is set to be high, the posterior probability of the middle purchase intention on game items becomes largest. This means that increase in visual authority value helps increase the purchase intention level. In addition, enjoyment value changes favorably most when the visual authority value increases. Therefore, game users seem to get more enjoyment value when their visual authority value of the game items grows significantly.

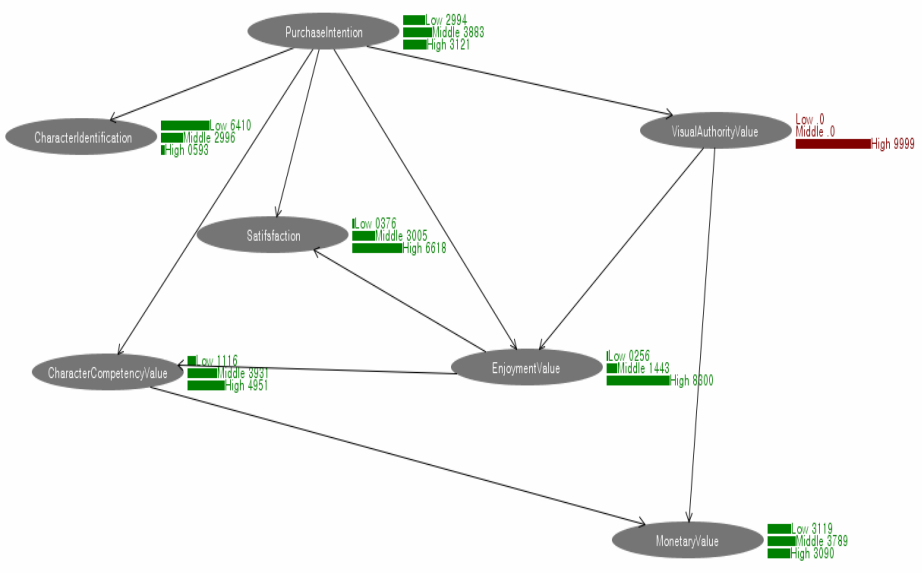


Fig. 2. What-if analysis result

Scenario 2 (Goal-Seeking analysis): For purchase intention to be high, what other factors should be changed?

Consequently, the purchase intention is affected by five variables: satisfaction, character identification, the enjoyment value, the visual authority value, and the character competency value. When the purchase intention is set to the highest level, the posterior probability for each high level of all of the variables increased. Among these variables, the levels of two variables, visual authority and character competency, changed. That is, the level with the highest posterior probability of visual authority changes from high to middle/high, and the level with the highest posterior probability of character competency changes from middle to high. However, in terms of posterior probability, satisfaction exhibits the largest increase, indicating that user's purchase intention is influenced significantly by visual authority value, character competency value, and satisfaction.

3.4 Discussion

From analyzing the GBN results, it was found that those factors affecting the purchase intentions are enjoyment value, visual authority value, character competency value, character identification, and satisfaction. In addition, visual authority value is related to enjoyment value, monetary value. Enjoyment value is associated with character competency value and satisfaction. Monetary value is related to character competency value and visual authority value. In addition, we found that most important node affecting the purchase intention is the character identification. There are several implications for the game companies.

First of all, it seems noteworthy that game companies should produce online games such that they are enjoyable and satisfactory to users. Users will purchase game items when the game is pleasant. Also to increase users' satisfaction level, game companies are asked to update games periodically and respond to users' complaint quickly.

Second, GBN results tell that the enjoyment value from buying game items is related to character competency value and satisfaction. In other words, users feel that they made a good bargain when their purchased game items can significantly add more power to their game characters than they thought when they purchased the game items. Therefore, game items should be designed to boost up the game character's power significantly once users pay considerably high price for the game items.

Third, GBN results show that game users do not seriously pay attention to the monetary value when their perceived satisfaction from buying the game items is greater than before the purchase of the game items. Therefore, game companies should always remain sensitive to users' need and satisfaction level.

Fourth, from the sensitivity analysis results, the importance of character identification was reassured because it is the most seemingly influential factor on user's perceived intention to purchase game items. Therefore, the higher character identification grows, the greater user's purchase intention becomes. Therefore, for enticing game users to buy online game items much more, it seems imperative for game companies to make game characters appear similar to a real person who game users might admire. Through the use of game characters similar to real persons, game users can get immersed better in game characters.

Finally, we proved that GBN is suitable for analyzing the causal relationship among related factors. In addition, by what if analysis and goal-seeking, more detailed analysis which is difficult in traditional methods like regression analysis is possible.

4 Concluding Remarks

The online gaming industry is growing as high speed Internet becomes more available in our daily life. Among the popularity of online games, the free-to-play games in which users are required to purchase game items to continue playing the games with more fun and more powerful game characters took sound position in the gaming industry. However, there is no study investigating the research issue of why game users are interested in purchasing the game items. Moreover, given several factors affecting user's intention to purchase game items, there is no study to attempt the analysis of causal relationships among them and propose practical guidelines for the game companies to come up with more effective game items-related strategy. In this respect, we proposed using the GBN to obtain a set of causal relationships and build useful management strategies about game item by performing a variety of sensitivity analyses.

Findings from using the GBN are as follows. First, we found that user's intention to purchase the game items are closely related to enjoyment value, visual authority value, character competency value, character identification, and satisfaction. Second, the visual authority value is associated with the enjoyment value and monetary value. Enjoyment value is related to character competency value and satisfaction, and monetary value is associated with the character competency value and the visual authority value. Third, it was revealed that the seemingly most important node affecting the

user's intention to purchase game items is the character identification. Fourth, through the sensitivity analysis using causal relationships supported by GBN, we found that the change in visual authority value influences purchase intention and enjoyment value significantly. Besides, it is required that the three values of visual authority, character competency, and satisfaction should remain high to sustain high level of user's intention to purchase game items.

To the best of our knowledge, this study is the first attempt that applies the GBN to resolve the research question- why users are interested in purchasing online game items. Through our experience with using the GBN to answer the research question, we came to firmly believe that GBN has great potentials enough to provide causal relationships as well as sensitivity analyses capability, all of which could be extremely useful for game companies to understand which factors are probably crucial in determining user's intention to purchase game items and accordingly design more appealing game items and suggest more effective management strategies in order to satisfy game users. Derivation of the GBN-based game items strategies like this would lead to mutually beneficial outcomes for both game companies and users.

Nevertheless, future study issues still remain unanswered. For example, a number of control variables including gender, age, gaming efficacy level, and disposable income need to be introduced. Also, the structural equation modeling approach needs to be applied to induce statistically significant results.

Acknowledgments. This study was supported by WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Hamari, J., Lehdonvirta, V.: Game Design as Marketing: How Game Mechanics Create Demand for Virtual Goods. International Journal of Business Science and Applied Management 5, 14–29 (2010)
2. Chiou, W.B., Wan, C.S.: A Further Investigation on the Motives of Online Games Addiction. In: National Educational Computing Conference, San Diego, U.S.A. (2006)
3. Korea Internet & Security Agency. Internet Usage Status Report in Korea (2009)
4. Lin, H., Sun, C.T.: Cash Trade within the Magic Circle: Free-to-Play Game Challenges and Massively Multiplayer Online Game Player Responses. In: DiGRA 2007, pp. 335–343 (2007)
5. Oh, G., Ryu, T.: Game Design on Item-Selling Based Payment Model in Korean Online Games. In: DiGRA 2007, pp. 650–657 (2007)
6. Cheng, J., Greiner, R.: Learning Bayesian Belief Network Classifiers: Algorithms and System. In: 14th Canadian Conference on Artificial Intelligence, pp. 141–151 (2001)
7. Madden, M.G.: On the Classification Performance of TAN and General Bayesian Networks. Knowledge-Based Systems 22, 489–495 (2009)
8. Guo, Y., Barnes, S.: Virtual Item Purchase Behavior in Virtual Worlds: an Exploratory Investigation. Electronic Commerce Research 9, 77–96 (2009)
9. Sheth, J.N., Newman, B.I., Gross, B.L.: Why We Buy What We Buy: a Theory of Consumption Values. Journal of Business Research 22, 159–170 (1991)
10. Long, M.M., Schiffman, L.G.: Consumption Values and Relationships: Segmenting the Market for Frequency Programs. Journal of Consumer Marketing 17, 214–232 (2000)

11. Park, H.J., Rabolt, N.J.: Cultural Value, Consumption Value, and Global Brand Image: a Cross-National Study. Psychology & Marketing 26, 714–735 (2009)
12. Finch, J.E.: The Impact of Personal Consumption Values and Beliefs on Organic Food Purchase Behavior. Journal of Food Products Marketing 11, 63–76 (2006)
13. Pope, N.: Consumption Values, Sponsorship Awareness, Brand and Product Use. Journal of Product & Brand Management 7, 124–136 (1998)
14. Williams, P., Soutar, G.: Customer Value and Tourism Satisfaction: a Multidimensional Perspective. In: ANZMAC 2005 Conference: Tourism Marketing, Fremantle, pp. 129–138 (2005)
15. Bødker, M., Gimpel, G., Hedman, J.: The User Experience of Smart Phones: a Consumption Values Approach. In: 8th global mobility roundtable, GMR 2009 (2009)
16. Lee, Y., Kim, J., Lee, I., Kim, H.: A Cross-Cultural Study on the Value Structure of Mobile Internet Usage: Comparison between Korea and Japan. Journal of Electronic Commerce Research 3, 227–239 (2002)
17. Turel, O., Serenko, A., Bontis, N.: User Acceptance of Hedonic Digital Artifacts: a Theory of Consumption Values Perspective. Information & Management 47, 53–59 (2010)
18. Live Gamer.: Virtual Item Monetization: A Powerful Revenue Opportunity for Online Game Publishers and Virtual World Operators (2008), http://www.livegamer.com/strategy/white-papers/Live_Gamer_Opportunity_Whitepaper_NA.PDF
19. Lehdonvirta, V.: Virtual Item Sales as a Revenue Model: Identifying Attributes that Drive Purchase Decisions. Electronic Commerce Research 9, 97–113 (2009)
20. Hefner, D., Klimmt, C., Vorderer, P.: Identification with the Player Character as Determinant of Video Game Enjoyment. In: Ma, L., Rauterberg, M., Nakatsu, R. (eds.) ICEC 2007. LNCS, vol. 4740, pp. 39–48. Springer, Heidelberg (2007)
21. McDonald, D.G., Kim, H.: When I Die, I Feel Small: Electronic Game Characters and the Social Self. Journal of Broadcasting & Electronic Media 45, 241–258 (2001)
22. Homburg, C.H., Rudolph, B.: Customer Satisfaction in Industrial Markets: Dimensional and Multiple Role Issues. Journal of Business Research 52, 15–33 (2001)
23. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers, San Mateo (1988)
24. Jensen, F.V.: Bayesian Networks and Decision Graphs. Springer, New York (2001)
25. Cooper, G.F., Herskovits, E.: A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning 9, 309–347 (1992)
26. Chan, H., Darwiche, A.: Reasoning about Bayesian Network Classifiers. In: Meek, C., Kjærulff, U. (eds.) 19th Conference in Uncertainty in Artificial Intelligence, pp. 107–115 (2003)
27. Cheng, J., Bell, D.A., Liu, W.: Learning Belief Networks from Data: An Information Theory Based Approach. In: 6th ACM International Conference on Information and Knowledge Management, pp. 325–331 (1997)
28. Cohen, J.: Defining identification: A Theoretical Look at the Identification of Audiences with Media Characters. Mass Communication & Society 4, 245–264 (2001)

U-BASE: General Bayesian Network-Driven Context Prediction for Decision Support

Kun Chang Lee[1], Heeryon Cho[2], and Sunyoung Lee[3]

[1] Professor of MIS and WCU Professor of Creativity Science
SKK Business School and Department of Interaction Science
Sungkyunkwan University
Seoul 110-745, Republic of Korea
kunchanglee@gmail.com, leekc@skku.edu
[2,3] Department of Interaction Science, Sungkyunkwan University
Seoul 110-745, Republic of Korea
heeryon@skku.edu, sunyoung92@gmail.com

Abstract. We propose a new type of ubiquitous decision support system that is powered by a General Bayesian Network (GBN). Because complicated decision support problems are plagued by complexities when interpreting causal relationships among decision variables, GBNs have shown excellent decision support competence because of their flexible structure, which allows them to extract appropriate and robust causal relationships among target variables and related explanatory variables. The potential of GBNs, however, has not been sufficiently investigated in the field of ubiquitous decision support. Hence, we propose a new type of ubiquitous decision support mechanism called U-BASE, which uses a GBN for context prediction in order to improve decision support. To illustrate the validity of the proposed decision support mechanism, we collected a set of contextual data from college students and applied U-BASE to induce useful and robust results. The practical implications are fully discussed, and issues for future studies are suggested.

Keywords: Context Prediction, General Bayesian Network, U-BASE.

1 Introduction

Context awareness has played an important role in enabling ubiquitous systems to serve as intelligent decision support systems [1, 2]. Such context awareness is based on the interpretation of contexts to understand in what kind of situations users are placed. Context is any information that can be used to characterize the situation of an entity, where an entity can be a person, place, or object that is considered relevant to the interaction between a user and an application [3]. When combined with ubiquitous computing systems [4], context awareness enables novel applications and services to adapt to a user's situation. Simple context awareness, however, does not guarantee proactiveness, which reduces a user's required efforts by predicting changes in relevant contexts for the future [5]. In other words, enabling ubiquitous decision support systems to be embedded with such proactiveness requires information about users' future needs, which must be inferred from users' future contexts. Predicting users'

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 63–72, 2010.

future context, which is called context prediction (CP), requires highly sophisticated inference methods capable of analyzing the given contextual data and finding meaningful patterns from them to predict future changes in user contexts.

Most CP problems pertain to location prediction [6] and action prediction [7]. When future locations that users are likely to visit soon (e.g., one hour later) are predicted precisely, a ubiquitous decision support system (UDSS) can provide timely and accurate decision support. Likewise, the UDSS will be accepted very favorably when the types of actions that decision makers take in the future are accurately forecasted. The literature has introduced various approaches as CP methods. For example, Laasonen et al. [8] define a hierarchy of locations and describe various methods that use statistics to predict a user's future locations. Patterson et al. [9] use a dynamic Bayesian network to predict likely travel destinations on a city map. Mozer et al. [10] use neural networks to predict how long a user will stay home and whether a particular zone will become occupied. Kaowthumrong et al. [11] use Markovian models to predict which remote control interface a user will likely use next. Petzold et al. [12] use global and local state predictors to predict the next room that a user will likely enter in an office environment. A more extensive methodological comparison was conducted by Mayrhofer [13], who compared the performances of different methods such as neural networks, Markov models, autoregressive moving average model (ARMA) forecasting, and support vector regression.

Though each CP method has a unique advantage over the others, all the methods have many pitfalls. The primary disadvantage is that most CP methods cannot establish causal relationships among the target variable and related explanatory variables. If such a causal relationship is extracted from target contextual data, it can be used to conduct a wide variety of what-if analyses. What-if analysis allows decision makers to see the possible results by varying the input conditions. In this way, the causal relationships obtained from the training dataset can be used as an inference engine that can perform various what-if analyses given the scenarios under consideration. To take advantage of the what-if analysis capability, we propose using a General Bayesian Network (GBN) in CP so that the causal relationships are induced from the training dataset, and future contexts can be inferred via what-if analyses for various scenarios. To illustrate the usefulness of GBN-powered CP, a system called Ubiquitous Bayesian network-Assisted Support Engine (U-BASE) is proposed, in which a GBN structure is used as a knowledge base to store a number of causal relationships among interested variables, and an inference engine is based on the what-if functions assisted by the GBN inference mechanism.

Hereafter, we explain the U-BASE design and usage scenario in Section 2 and experiments using real contextual dataset in Section 3. Section 4 presents a discussion of the implications of the GBN-powered CP, and Section 5 offers concluding remarks and suggestions for future research issues.

2 U-BASE

The U-BASE system collects user transaction data to construct BN models and predicts a user's future contexts using the BN models to provide context-sensitive recommendations to users. Figure 1 shows the U-BASE system architecture.

2.1 Design

The U-BASE system consists of five components (a data collection component, a BN model learning component, a BN model registration component, a context prediction component, and a recommendation component), a BN model base, and a set of databases that store both context and the factual data (Fig. 1).

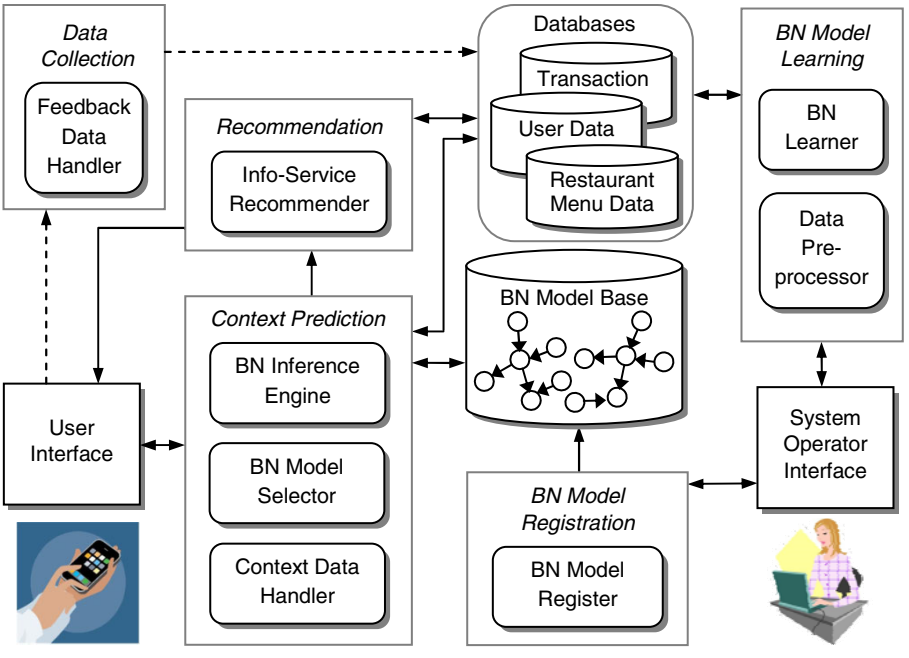


Fig. 1. U-BASE system architecture

Data Collection Component. The data collection component collects user feedback sent from user devices and stores feedback information to the transaction database. The feedback information acts as a class label to determine whether the recommendations (or future contexts) given by the system are useful and correct. The user transaction data coupled with the user feedback information are used later as training and test data for the induction of BN models.

BN Model Learning Component, BN Model Registration Component, and BN Model Base. BN models can be built using two different approaches – the data-based approach or the knowledge-based approach. The former approach induces BN models from user history data, whereas the latter approach manually constructs BN models by employing the domain knowledge of human experts [14]. Once a good BN model is constructed, either using a BN model learning component or by a human expert, the BN model registration component registers it to the BN model base. More specifically, the constituents of each component are as follows:

Data Preprocessor: The data preprocessor retrieves user history data from the transaction database to create the training/test data needed to create the BN model. It interacts with humans via the system operator interface to preprocess training/test data.

BN Learner: The BN learner creates BN models from the training data, but human intervention is required to set the target variable and parameters for learning. New BN models are learned as new recommendation services are added to the system. The prediction accuracy of the BN model is checked to determine the model's quality. The iterative process of adjusting the parameters and checking the prediction accuracy is repeated until a good BN model is built.

BN Model Register: Once a satisfactory BN model is learned, the BN model register registers the BN model to the BN model base. Manually constructed BN models are also registered to the BN model base via the BN model register.

BN Model Base: The BN model base maintains multiple BN models for different context prediction-based services. These BN models predict future contexts such as the next location, next activity, and next goal, among other things.

Context Prediction Component. The key component of the U-BASE system is the context prediction (CP) component. The CP component consists of a context data handler, a BN model selector, and a BN inference engine. The context data handler passes user context data to the BN model selector, and the BN model selector selects an appropriate BN model from the BN model base on the basis of user context data. The BN inference engine then performs context prediction on the basis of the selected BN model and the context data and passes the predicted results back to the context data handler. The context data handler then passes the results to the recommendation component.

Context Data Handler: The context data handler receives user context data from user applications in two ways: it can receive context data that are deliberately sent by the user (user-initiated), or it can receive data by proactively requesting the user application for context data (system-initiated). In some cases, not all the context data will be available via user applications. In such cases, additional context data may be obtained from the databases. For example, the user application may pass only the user ID to the context data handler, and the rest of the user data may be retrieved from the user database.

BN Model Selector: The BN model selector selects an appropriate BN model from the BN model base on the basis of the user context data, and then passes the selected BN model and user context data to the BN inference engine for context prediction.

BN Inference Engine: The BN inference engine performs what-if simulation on the selected BN model using user context data. The target variable's entries' posterior probabilities are calculated by instantiating the explanatory variables; the entry with the greatest probability is given as the predicted result.

Recommendation Component. The info-service recommender utilizes user context data, the predicted results, and relevant factual data retrieved from databases to generate context-sensitive information useful to the user. The final information may be filtered or edited to further match a user's needs and expectations.

Databases. A set of databases stores and maintains both the context and factual data for the U-BASE system. These data are broadly classified into three categories. The first data category deals with the user context data required for predicting the future context. These data are primarily gathered from the user application, but additional context data may be retrieved from databases to supplement the context data.

The second data category deals with the factual data used for generating the final information presented to the user. These factual data constitute the ingredients of the final output information and are often stored and updated in distributed databases.

The third data category deals with transaction data used for learning BN models. These data are in most cases collected and generated from the first and second category data, that is, user context data and factual data, or a combination of the two with additional user feedback. Each user transaction record can be considered the training/test instances for building BN models for context prediction. The U-BASE system continuously manages and updates these diverse and voluminous data through multiple, distributed databases.

2.2 Usage Scenario

We now present a scenario that demonstrates how a GBN is used for context prediction in the U-BASE system. Imagine a smart phone service targeted toward college students to assist their daily activities on campus. One campus information service includes a food menu recommendation service, which predicts a user's future location (i.e., the place a user will visit next) to provide nearby restaurant recommendations for the future location. It is eleven in the morning, and as Tom, a sophomore majoring in social science, is contemplating what to eat for lunch, he receives a message from the U-BASE system that asks for his current location. The student sends his current location information ('Suseon Hall') to the service, and the context data handler inside the context prediction component receives the current location data and retrieves additional user data such as the student's major and year in school from the user's database using his user ID. (Refer to Fig. 1 as needed.) The user context data (student major, student year, current location, and future activity) are then sent to the BN model selector, and the BN model selector picks out a relevant BN model from the BN model base according to the user context data. In this case, the BN model selector picks out a BN model that predicts the user's next location (Fig. 2).

The selected model and user context data are then passed onto the BN inference engine, and a what-if simulation is performed by setting the 'Activity' node's evidence to 'Eat', the 'Major' node to 'Social Science', the 'Year' node to 'Sophomore', and the 'Location Departed' node to 'Suseon Hall'. Consequently, the target node ('Location Arrived') entries are influenced by this instantiation and give '600th Anniversary Building' the largest posterior probability, making it the next location value. The context data handler receives this next-location value ('600th Anniversary Building') from the BN inference engine and passes the next-location value and user context data to the recommendation component. The recommendation component retrieves today's menu served at the '600th Anniversary Building' cafeteria from the restaurant menu database and sends it as the final output to the user's application.

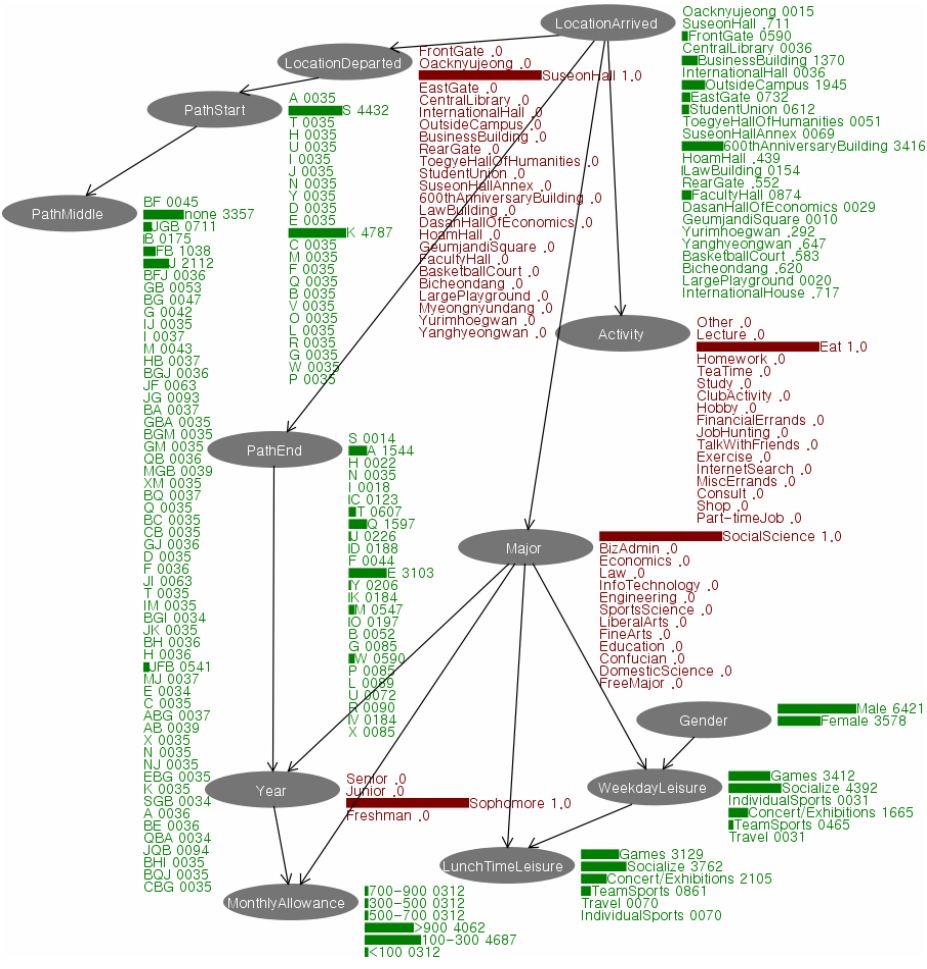


Fig. 2. A general Bayesian network for next-location prediction with 'Location Arrived' as the target node

3 Experiment

The heart of the U-BASE system lies in its context prediction component; the system would be far less useful if the predicted-context accuracy was low. In this section, we investigate how effective GBN is compared to that of other Bayesian network classifiers. To do this, we collect contextual data from undergraduate students to construct the Bayesian networks (BN models) mentioned in the previous section (Fig. 2). We build three types of Bayesian networks, a General Bayesian Network (GBN), a Naïve Bayesian Network (NBN), and a Tree-Augmented Naïve Bayesian Network (TAN), and compare their prediction accuracies.

3.1 Data and Variables

Campus activity data were collected from college students to create user context data for the experiment. The college students were shown a campus map containing building and route information, as depicted in Fig. 3, and were asked to document their two days of activities on campus. They documented where they visited via what route (a list of letters in Fig. 1 was specified to describe a sequence of paths) and what activity they engaged in at that location. To describe the activity, they chose one of seventeen predefined activity values listed in the 'Activity' node in Fig. 2.

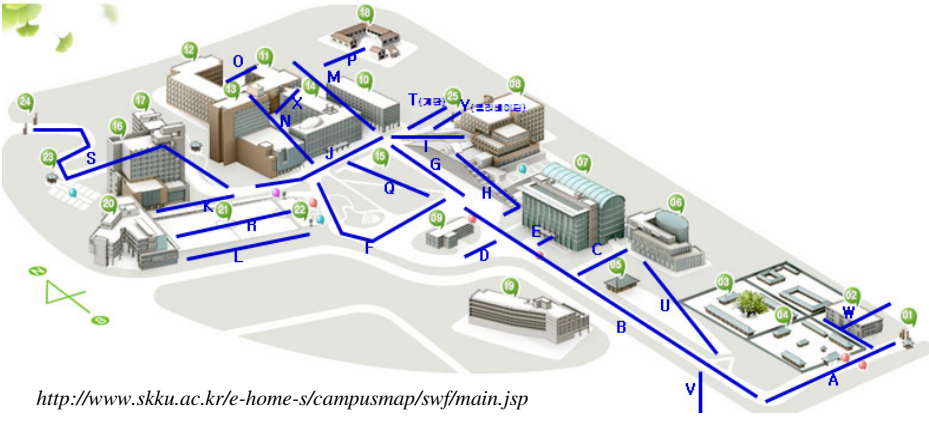


Fig. 3. A campus map containing building and route information

In addition to campus activity data, the students filled out a questionnaire that asked their gender, major, student year, weekday leisure activities, lunch-time leisure activities, monthly allowance, and student ID. The experiment used data from 335 students.

After all data were cleaned, campus activity data and personal data were combined to create campus activity-demographic data. Student IDs were used to help combine the two types of data. The combined data contained twelve attributes ('Location Arrived', 'Path Start', 'Path Middle', 'Path End', 'Location Departed', 'Activity', 'Gender', 'Major', 'Year', 'Weekday Leisure', 'Lunch Leisure', and 'Monthly Allowance'). A total of 3,150 records of campus activity-demographic data were used to construct three types of Bayesian networks, a GBN (Fig. 2), an NBN, and a TAN.

3.2 Structure Learning

We used WEKA [15], an open source data-mining tool with Bayesian network learning and inference capabilities, to construct the Bayesian networks and to perform the experiments. The twelve-variable campus activity-demographic data were used to create networks with 'Location Arrived' as the target node. The structure of the GBN was learned by using a K2 algorithm [16] with the maximum number of parents node limited to two. To construct the NBN and TAN [17], the default setting in WEKA was used. In all cases for the GBN and TAN, the BAYES scoring metric was used.

3.3 Results

Table 1 lists the accuracy (and standard deviation of accuracy) of the GBN, NBN, and TAN classification algorithms as measured from 10 runs of 10-fold cross-validation. The best of the three classifier results is displayed in bold and was determined by using a corrected resampled t-test [18] at the 1% significance level based on the 10 X 10 fold cross-validation results. The results show that there are statistical differences between the GBN and NBN and between the GBN and TAN, and that in both cases the GBN is better. Moreover, we performed one-way ANOVA using a significance level of $\alpha = .05$ to confirm that the mean difference between the three approaches is significantly different (F (2, 297) = 229.208, $p < .001$). *Post hoc* analyses using the Scheffe *post hoc* criterion for significance indicates that all three approaches' means are significantly different from one another ($p < .001$); specifically, the GBN is better than both the NBN and the TAN, and the NBN is better than the TAN. When we look at one specific run of the ten runs, we see that, given 3,150 test instances, the GBN makes 492 false predictions, whereas the NBN and TAN make 566 and 667 false predictions, respectively.

Table 1. Prediction performance (accuracy ± std dev) of three algorithms

Target Node	GBN	NBN	TAN
Location Arrived	**84.12 ±1.70**	81.90 ±1.79	78.87 ±1.73

4 Discussion

We confirmed the prediction accuracy of the GBN to outperform both performances of the NBN and TAN, but better performance alone does not make a GBN a good classifier for context prediction. The structure of a GBN is much more flexible than the fixed structure of the NBN and the TAN, endowing GBN with the capacity to express cause and effect between not only the target variable and the explanatory variables, but also among the explanatory variables themselves.

Greater representational power, however, does not necessarily mean greater complexity. The GBN may have fewer links than the TAN for the same domain, since not all nodes are linked to the target node [19]. This is also true for the GBN and the TAN presented in this paper; counting the links between the target node and the explanatory nodes, we see that the GBN has four links, whereas TAN has eleven.

Better prediction accuracy, greater representational power, and low complexity are all strengths of the GBN, but the greatest advantage of the GBN is that fewer variables are required for context prediction. As demonstrated in Fig. 2, the target node is directly linked to fewer explanatory nodes (four) than those of the NBN or the TAN (eleven for both approaches). Hence, the selectiveness of the GBN allows humans to grasp which explanatory nodes (variables) are crucial for target node prediction. Because obtaining data for instantiation sometimes can be costly and difficult, knowing which variables are more important than others can be efficient and effective, and this knowledge can help build a data collection strategy for better context prediction.

Context prediction can also improve human/computer interaction to provide better service that conforms to a user's expectations. Here, prediction accuracy is also crucial to the success of context prediction-based services. One way to achieve good context prediction accuracy is to use different BN models as necessary. For example, a first-time user may not have enough personal transaction data, so it may be difficult to create a BN model that adequately reflects the user. In such cases, the system can first construct a BN model by using the transaction data of a group of users sharing similar traits with those of the first-time user. In the beginning, the system can use the group BN model to predict contexts for the first-time user. Over time, as the first-time user's own transaction data gradually accumulate, the system can create a new BN model that better models the user's profile and use it for context prediction.

5 Concluding Remarks

The main contributions of this paper are: (a) an evaluation of the performance of the U-BASE through the use of real-world contextual datasets to determine its effectiveness for resolving context prediction (CP) problems, and (b) a demonstration that GBN-assisted ubiquitous decision support for CP is efficient and robust in real-world situations.

As to the first contribution, the statistical test results summarized in Table 1 show that the GBN-assisted UDSS performs best in comparison with the other two BNs, the NBN and the TAN. At the 99% confidence level, the GBN-assisted UDSS performed best for the given target node. As to the second contribution, the GBN-based inference mechanism for resolving CP problems was shown to be useful in a situation in which there are many variables to be considered, and the target node seems to depend causally on many explanatory variables. Since a GBN provides a set of causal relationships among the variables under consideration, the causal relationships given by a GBN can be stored into the knowledge base on the basis of which various types of what-if simulations can be performed to induce CP solutions for the target users.

Future research directions include a user evaluation of the U-BASE system and further comparison of the GBN-assisted inference mechanism with other inference methods such as neural networks and decision trees, among others. Moreover, the improvement of prediction performance through ensemble methods, which combine multiple classifiers such as neural network and decision trees, should be studied to produce more robust and more accurate context prediction.

Acknowledgments. This research was supported by the WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Cook, D.J., Augusto, J.C., Jakkula, V.R.: Ambient Intelligence: Technologies, Applications, and Opportunities. Pervasive and Mobile Computing 5, 277–298 (2009)
2. Shim, J.P., Warkentin, M., Courtney, J.F., Power, D.J., Sharda, R., Carlsson, C.: Past, Present, and Future of Decision Support Technology. Decision Support Systems 33, 111–126 (2002)

3. Dey, A.K., Abowd, G.D.: Towards a Better Understanding of Context and Context-Awareness. In: Gellersen, H.-W. (ed.) HUC 1999. LNCS, vol. 1707, pp. 304–307. Springer, Heidelberg (1999)
4. Weiser, M.: The Computer for the 21st Century. Scientific American 272, 78–89 (1995)
5. Tennenhouse, D.: Proactive Computing. Communications of the ACM 43, 43–50 (2000)
6. Petzold, J., Bagci, F., Trumler, W., Ungerer, T.: Next Location Prediction within a Smart Office Building. Cognitive Science Research Paper, University of Sussex CSRP 577, 69 (2005)
7. Kim, E., Helal, S., Cook, D.: Human Activity Recognition and Pattern Discovery. IEEE Pervasive Computing 9, 48–53 (2010)
8. Laasonen, K., Raento, M., Toivonen, H.: Adaptive On-Device Location Recognition. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 287–304. Springer, Heidelberg (2004)
9. Patterson, D., Liao, L., Fox, D., Kautz, H.: Inferring High-Level Behavior from Low-Level Sensors. In: Dey, A.K., Schmidt, A., McCarthy, J.F. (eds.) UbiComp 2003. LNCS, vol. 2864, pp. 73–89. Springer, Heidelberg (2003)
10. Mozer, M.C.: The Neural Network House: An Environment that Adapts to Its Inhabitants. In: American Association for Artificial Intelligence Spring Symposium on Intelligent Environments, pp. 110–114 (1998)
11. Kaowthumrong, K., Lebsack, J., Han, R.: Automated Selection of the Active Device in Interactive Multi-Device Smart Spaces. In: Workshop at UbiComp 2002: Supporting Spontaneous Interaction in Ubiquitous Computing Settings (2002)
12. Petzold, J., Bagci, F., Trumler, W., Ungerer, T., Vintan, L.: Global State Context Prediction Techniques Applied to a Smart Office Building. In: The Communication Networks and Distributed Systems Modeling and Simulation Conference (2004)
13. Mayrhofer, R. (ed.): An Architecture for Context Prediction, Advances in Pervasive Computing, part of the 2nd International Conference on Pervasive Computing, vol. 176, pp. 65–72. Austrian Computer Society (OCG) (2004)
14. Nadkarni, S., Shenoy, P.P.: A Causal Mapping Approach to Constructing Bayesian Networks. Decision Support Systems 38, 259–281 (2004)
15. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter 11, 10–18 (2009)
16. Cooper, G.F., Herskovits, E.: A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning 9, 309–347 (1992)
17. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers. Machine Learning 29, 131–163 (1997)
18. Nadeau, C., Bengio, Y.: Inference for the Generalization Error. Machine Learning 52, 239–281 (2003)
19. Madden, M.G.: On the Classification Performance of TAN and General Bayesian Networks. Knowledge-Based Systems 22, 489–495 (2009)

A Dynamic Bayesian Network Approach to Location Prediction in Ubiquitous Computing Environments

Sunyoung Lee[1], Kun Chang Lee[2], and Heeryon Cho[3]

[1,3] Department of Interaction Science, Sungkyunkwan University
Seoul 110-745, Republic of Korea
sunyoung92@gmail.com, heeryon@skku.edu
[2] Professor of MIS and WCU Professor of Creativity Science
SKK Business School and Department of Interaction Science
Sungkyunkwan University
Seoul 110-745, Republic of Korea
kunchanglee@gmail.com, leekc@skku.edu

Abstract. The ability to predict the future contexts of users significantly improves service quality and user satisfaction in ubiquitous computing environments. Location prediction is particularly useful because ubiquitous computing environments can dynamically adapt their behaviors according to a user's future location. In this paper, we present an inductive approach to recognizing a user's location by establishing a dynamic Bayesian network model. The dynamic Bayesian network model has been evaluated with a set of contextual data collected from undergraduate students. The evaluation result suggests that a dynamic Bayesian network model offers significant predictive power.

Keywords: Dynamic Bayesian Networks, Context Prediction, Ubiquitous Computing.

1 Introduction

In recent years, context-awareness has been the subject of growing attention in the area of ubiquitous computing because of its usefulness in several different applications [5]. When computer systems are aware of the context in which they are used and can adapt to changes in context, they can engage in more efficient interaction with users. Context awareness involves enabling ubiquitous computing devices to be aware of changes in the environment and to intelligently adapt themselves to provide more meaningful and timely decision support for decision-makers [4]. However, context-aware systems are limited by the fact that their target is the current context and that context-aware systems do not predict the future context. Therefore, the qualities of services provided by context-aware systems are seriously restricted when future contexts change drastically. To this end, we need to consider the task of context prediction in order to proactively offer high-quality services for users in ubiquitous computing environments.

Context prediction opens a wide range of possibilities for context-aware computing applications. A context-prediction application may infer the future location of an office owner and redirect incoming calls to that future location. A context-prediction application may also be useful for enhancing the qualities of transportation systems.

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 73–82, 2010.

On the basis of information about the current and future locations of a particular user, transportation systems equipped with context-prediction technology may be able to more effectively assist drivers by inferring possible preferred routes and by providing customized route suggestions for drivers, as well as by warning drivers about possible dangers by predicting their future context. Knowing the current location and current time, together with the user's calendar, could also allow applications to have a good idea of the user's current social situation, such as if the user is in a meeting, in class, or waiting in the airport.

The list of applications listed here is limited, and we believe that there is great potential for context prediction to be used in a variety of ubiquitous computing applications. This paper proposes a Dynamic Bayesian Network (DBN) approach to location prediction for ubiquitous computing environments. A DBN is an important tool because it can represent the temporal properties of user-context information. In fact, it is obvious that a user's current locations are influenced by their previous locations, and that particular locations are related to particular types of actions. Therefore, we adopted a DBN approach for recognizing user locations.

This paper is structured as follows. Section 2 discusses context prediction and various context prediction techniques in ubiquitous computing environments. The modeling techniques used to predict a user' locations are described in Section 3. The results of the experiment are presented and discussed in Section 4, followed by concluding remarks and directions for future work in Section 5.

2 Background

2.1 Context Prediction

Context prediction attempts to infer user context by analyzing observed context history, a series of contextual information data that show how users are moving around in a certain ubiquitous computing environment. The contextual information is supplied by various types of sensors such as GPS, RFID, and various wireless devices that may provide contextual information about users' locations, actions, or changes in physical environment. The purpose of context prediction is to predict the subsequent context that users will likely enter (if the contexts are locations or situations) or perform (if the contexts are actions) on the basis of a history of contexts, which are compiled through various sensors. For ubiquitous computing environments, the ability to accurately predict a user's contexts would make it possible to provide context-aware services that are more natural and customized to people's needs. Accurately recognizing a user's contexts can provide more effective and personalized guidance to a user, particularly in ubiquitous decision support systems.

2.2 Context-Prediction Techniques

Several context-prediction techniques have been proposed in the literature, including Bayesian networks [8], [14], Markov models [15], [16], topic models [7], [9] and neural network approaches [14]. Examples of context prediction are location prediction [1], [10], [14], movement prediction [13], action prediction [2], [3], [16], and daily routine prediction [7], [9].

Petzold *et al*. [14] investigated Bayesian networks, neural networks, Markov models and state predictors to predict the next location of an office owner in an office building. Singla *et al*. [16] proposed a Hidden Markov model approach for recognizing activities performed by multiple residents in a single smart-home environment. A Hidden Markov model was used to determine an activity that most likely corresponds to an observed sequence of sensor readings. Bayesian networks can be used to predict the prominent activities of users. For example, Hwang and Cho [8] proposed a modular Bayesian network model to infer the landmarks of users from mobile log data such as GPS logs, call logs, SMS logs, picture logs, music-playing logs, and weather logs.

A promising topic model approach for recognizing a user's daily routines has been proposed for ubiquitous computing environments. For example, Huýnh *et al*. [7] adopted a topic model to predict users' daily routines (such as office work, commuting, or lunch routine) from activity patterns. To evaluate the topic model, they collected the daily activities of one person over a period of 16 days. For data collection, the subject wore two sensors in order to record low-level signals such as body movements or body posture. The subject was asked to annotate his activities in detail in order to model the relationship between user activities and low-level signals. In total, 34 activities were recorded in their dataset. Huýnh *et al*. first identified the user's activity patterns from low-level sensor data using various classifiers such as Support Vector Machines, Hidden Markov models, and a naïve Bayesian network. The resulting user activity patterns that were identified were then used in the topic model as inputs to infer the user's daily routine.

As we have seen so far, many examples of context prediction exist in a variety of application domains. Several strategies can be employed to identify the future location of a user, including probabilistic models that predict a user's future location. Section 3 describes an inductive approach for generating location prediction models in ubiquitous computing environments.

3 Inducing Location Prediction Models

Many types of location recognition models can be used. We investigated probabilistic models such as Dynamic Bayesian Networks (DBNs), General Bayesian Networks (GBNs), Tree Augmented Naïve Bayesian networks (TANs), and Naïve Bayesian networks (NBNs).

3.1 Bayesian Models for Location Prediction

A Bayesian Network approach is well suited for generating predictive models in a real-world domain because it can deal with the uncertainty inherent in every facet of human life. Bayesian networks are probabilistic models in the form of directed acyclic graphs [12], in which nodes represent variables or propositions (e.g., the occurrence of an event or the feature of an object). Likewise, links represent causal or informational dependencies among variables and are quantified by the conditional probability of a node, given its parents. If a node does not have parents, it is associated with a prior probability. Since Bayesian networks represent causal or informational dependencies among variables, variables that are not influenced by any other

variables but do exert an influence on other variables are positioned in the top layer of the network. Similarly, variables that both are influenced by some variables and also influence other variables are positioned in the middle layers of a network, while variables that are influenced by some variables but that do not influence any other variables are positioned in the bottom layer. In such a representation, the probability of any combination of variables can be inferred without having to represent the joint probabilities of the variables.

In general, there are five classes of Bayesian networks. An NBN [17] is the simplest Bayesian network that has one node that is a parent to all of the other nodes. The parent node is often called a class node. No other links exist in the network. NBN is useful for preliminary predictive model induction because of its naïve independence assumptions. A TAN [17] is formed by adding directional links between attributes in an NBN. After removing the class node in a TAN, the attributes should form a tree. A GBN [17] is an unrestricted Bayesian network that treats the class node as an ordinary node. Therefore, the class node can be a child node of other nodes. These Bayesian networks do not provide a direct mechanism for representing temporal dependencies among attributes. However, most real-world events change over time, and such change cannot be modeled using static Bayesian networks. A DBN is, therefore, a Bayesian network that represents variables with temporal characteristics and is composed of a sequence of General Bayesian networks. A DBN [11] provides a systematic method for modeling the temporal and causal relationships among variables. In a DBN, each Bayesian network represents the states of variables at different times.

3.2 GBN, TAN, and NBN Model Induction

For model induction, we adopted two different learning techniques, those that are fully automated, and those that require knowledge provided by a domain expert. We adopted the completely automated approach to generate GBN, TAN, and NBN location prediction models. During the training phase, the following observable attributes were recorded:

- *Previous User Actions*: previous user action is action that a user took immediately preceding the current actions
- *Current User Actions*: current user action is action that a user is currently taking
- *User Locations*: user location is the location in which a current user action is being performed
- *Routes*: routes are the paths that a user can take to arrive at the current location. In our experiment, the maximum number of routes that a user could take was seven. For example, route 1 was the first route that a user could take from the previous location to get to route 2, and route 2 represented a second route that a user could take from route 1 to route 3. Therefore, routes 1 ~7 comprise location paths from a previous location to the current location.

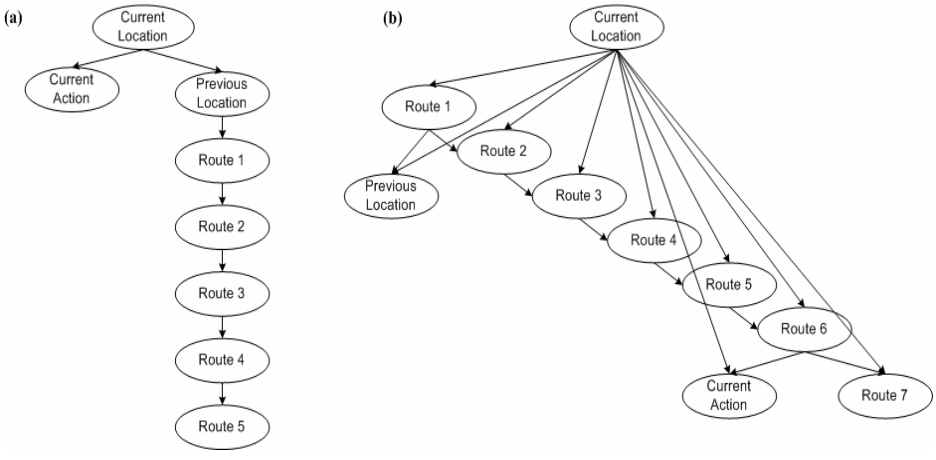


Fig. 1. GBN and TAN Location Prediction Models using a Weka machine learning tool. (a) Learned GBN model and (b) Learned TAN model.

Once the dataset was prepared, the user data were loaded into the WEKA machine learning tool [17], the structures and conditional probabilities of GBN, TAN, and NBN were learned, and ten-fold cross-validation analyses were run on the resulting models. The entire dataset was used to generate several types of location prediction models. Figure 1 shows the induced GBN and TAN location recognition models. Because most users did not record route 6 and 7, nodes representing these routes were not included in the induced GBN.

3.3 DBN Model Induction

The model induction for DBN proceeds in four phases: (1) identify domain variables; (2) examine dependencies among the domain variables and the manners in which these domain variables change over time; (3) describe how the conditional probability distributions are constructed from the user's actions and location data; and (4) procedurally develop the belief update. The domain variables that were used in the DBN model included high-level actions (such as talking, walking, and moving) and places (such as a classroom, a house, and outdoors) in which users could perform these actions. After indentifying the domain variables, their dependencies were determined.

The DBN shown in Fig. 2 represents the dependencies between actions and locations. Because current actions are affected by previous actions, there is a directional link from user action at time t-1 to user action at time t. This temporal link shows that the domain variables change over time. Further, links from user location nodes to user action nodes indicate that a user action is executed at some location.[1]

[1] Locations are modeled as influencing actions because particular types of actions executed at particular locations. However, the converse model in which actions influence locations is also plausible because a user might move to a location to perform an intended action there. Because Bayesian networks are by definition acyclic, one of these directions of causality must be selected. The former is chosen here.

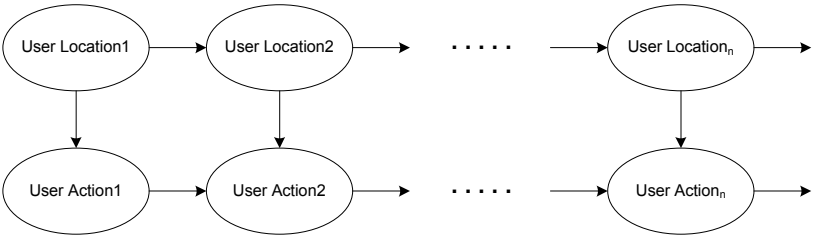


Fig. 2. DBN Location Prediction Model

Once the structure of the DBN is specified, P(User Action t | User Action t-1), P(User Action t | User Location t), and P(User Location t | User Location t-1) are estimated using the dataset in the training phase. Because not all possible configurations of variables can typically be observed, the conditional probabilities of some variables cannot be computed easily from the data. For example, if a network has N nodes (variables), and even if each node can have discrete binary values, the total number of possible configurations is 2^N. Therefore, it is often the case that data are insufficient for learning the conditional probabilities. In order to avoid zero probabilities, a small number is often added to each cell in sparse conditional probability tables [6]. Often called a flattening constant (denoted by α), this number can be added either only to empty cells or to all cells in the table. After adding α to the cells, the conditional probabilities are recomputed. While different choices of α have been proposed (e.g., adding 1/2 to all cells or adding 1/D to empty cells where D is the total number of cells), we chose to reevaluate the conditional probabilities after adding 1 to all cells in the conditional probability tables for the DBN location prediction model.

In the testing phase, the DBN location prediction model begins to predict the user's location at time 1 by creating slice 1. In slice 1, only two nodes exist, User Action 1 and User Location 1, all of the available information at the time point. Therefore, we set the current action as evidence for the node User Action 1 and compute the posterior probability P(User Location 1 | User Action 1= observed current action) by using a belief update algorithm, for which any belief update algorithms can be used to propagate beliefs through the DBN. The DBN location prediction model subsequently determines the location with the maximum posterior probability value as the most probable user location at time 1. At time 2, the DBN location prediction model creates another slice for time 2. At this point in time, the DBN location prediction model consists of two slices for times 1 and 2. The current evidence is the user action and location at time 1 and the user action at time 2. From this information, the DBN prediction model makes a prediction about the user location at time 2 by considering this information evidence and computing the posterior probability P(User Location 2 | User Location 1= the observed location at time 1, User Action 1 = the observed action at time 1, User Action 2 = the observed action at time 2). The location with the maximum posterior probability value now becomes the most probable user location at time 2. We describe this inference process below.

For i = 1, … , number of folds:

1. Partition the dataset into training set i and test set i
2. Estimate the following prior and conditional probabilities using training set i:

   $P(L_1)$, $P(L_t \mid L_{t-1})$, $P(A_t \mid L_t)$, $P(A_t \mid A_{t-1})$,

   where A_t and L_t represent user action and location at time t, respectively.
3. For j = 1,…, number of time slices:

   3.1 Create a slice for time j

   3.2 Set the evidence that has been observed so far:

   If j ==1 then:

   $A_1 \leftarrow a_1$

   Else:

   $A_1 \leftarrow a_1 \quad L_1 \leftarrow l_1$

   … …

   $A_j \leftarrow a_j \quad L_{j-1} \leftarrow l_{j-1}$

   3.3 Compute the following posterior probability using any DBN belief update algorithm:

   $P(L_j \mid l_{1:j-1}, a_{1:j})$, where the action sequence a_1, a_2, ..., a_j is denoted by $a_{1:j}$ and the location sequence l_1, l_2, …, l_{j-1} is denoted by $l_{1:j-1}$.

   3.4 Choose the most probable action L* as follows:

   $L^* \leftarrow \text{argmax} P(L_j \mid l_{1:j-1}, a_{1:j})$
4. Output the predicted location sequence.

4 Evaluation

In a formal evaluation, data were gathered from 366 undergraduate students at a private university in Seoul, Korea. In order to motivate participation, 2% of the subject's total class points were given as extra credit points. There were 125 female and 211 male participants; the average age of the female subjects was 20.7 years old, and the average age of the male subjects was 22.6 years old.

After filling out a demographic survey, participants were asked to record their daily routines (e.g., what they are doing and where they are when they perform a particular action) on campus over a period of two days. Record-keeping began when the subjects arrived at school and ended when they left school for the day. In order to obtain the dataset, we employed a time diary, which in our case was a handwritten log in which the subject wrote the start and end times of each action, the location in which the action was performed, and the routes that the subject took in order to arrive at the location. Campus location codes, route codes, and action codes were provided to the subjects in order to help them record their daily routines. Initially, there were 84 distinct actions, 25 distinct routes, and more than 30 distinct locations that the subjects could choose from, of which our subjects recorded a total of 30 distinct locations. We obtained a total of 672 days worth of activity data. Daily activity data were removed if the length of time recorded was too short (e.g., only one action was

recorded) or if the subjects recorded a route that they could not have taken given their positions. As a result, 266 out of 672 days of activity data were used in the formal evaluation. The examples of actions, locations, and routes that users could take are described in Table 1.

Table 1. Examples of possible values of domain variables

Attribute	*Examples of Possible Values*
Action	Attending class/attending a seminar, preparing for exam/doing homework, doing things that are not listed, eating lunch, talking to friends, studying, eating snack/smoking/having tea time, doing club activities, surfing the Internet, working at a part-time job, buying things, meeting with a professor, etc.
Location	600^{th} anniversary building, basketball court, Bicheondang, business building, central library, Dasan hall of economics, east gate, faculty hall, front gate, Geumjandi square, Hoam hall, international hall, large playground, law building, Myeongnyundang, Oacknyujeong, outside campus, rear gate, student union building, Suseon hall, Suseon hall annex, Toegye hall of humanities, Yanghyeongwan, Yurimhoegwan, etc.
Route	A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W

4.1 Results

The DBN, GBN, TAN, and NBN were created from data collected using the method described above. The induced models were evaluated using a ten-fold cross-validation. For the DBN model, the structure was fixed; conditional probabilities were learned. For the GBN, TAN, and NBN models, the structure and conditional probabilities were determined. Table 2 reports the overall results of the DBN, GBN, TAN, and NBN location recognition models. The percentages refer to correctly classified instances. The highest-performing model was a DBN location recognition model (F-statistics = 29.65, $p < 0.00001$), and the second-highest performing model was a TAN location recognition model. The accuracies of the GBN and NBN were lower than those of the DBN and TAN models, but they performed significantly better than chance, which was 3.33% based on the presence of 30 candidate locations. The results suggest that the DBN location prediction model reported here can accurately identify user locations.

Table 2. Evaluation Results

Model	*DBN*	*GBN*	*TAN*	*NBN*
Average Accuracy (%)	72.67	45.88	69.29	55.27

This experiment has two important implications for the design of location prediction modeling in ubiquitous computing environments. First, by monitoring a sequence of user locations and actions in a ubiquitous computing environment, induced models can make accurate predictions about forthcoming user locations. Second, using

models that can predict user locations creates a significant window of opportunity within which to take corrective action in a ubiquitous environment; this is because context-prediction models improve on context-aware approaches that predict only the current contexts of users.

5 Conclusion and Future Work

Context prediction is an important problem in ubiquitous computing environments. Accurately predicting user contexts could greatly improve the quality of user satisfaction in every aspect of daily life, particularly in ubiquitous computing environments. By drawing inferences about user locations, ubiquitous systems not only can automatically detect a user's current situation, but also can forecast the user's likely future location. Such location prediction systems will help users make decisions quickly and efficiently by informing users as to the location of the most suitable services, given their situation and needs.

This paper introduces an inductive approach for generating location prediction models in ubiquitous computing environments. In this approach, models are created from observations of user interactions within a ubiquitous computing environment in which user actions, locations, and temporal information are monitored. After user interaction traces have been recorded, location prediction models are developed.

The findings reported in this study contribute to a growing body of work on context prediction for ubiquitous computing environments. In the future, it will be useful to investigate much richer contextual information such as a user's intention or emotional states on which to predict a user's future context. Second, evaluating the resulting models as a runtime component in ubiquitous environments will be an important next step in developing context-prediction applications. Third, it will be important to investigate location prediction models that can make "early" predictions of user locations. Early prediction will allow ubiquitous systems adequate time to prepare services for a user's particular future location or to suggest alternative locations before users move to dangerous locations.

Acknowledgments. This research was supported by the WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Anagnostopoulos, T., Anagnostopoulos, C., Hadjiefthymiades, S., Kyriakakos, M., Kalousis, A.: Predicting the Location of Mobile Users: a Machine Learning Approach. In: The ACM International Conference on Pervasive Services, pp. 65–72 (2009)
2. Brdiczka, O., Reignier, P., Crowley, J.: Detecting Individual Activities from Video in a Smart Home. In: 11th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pp. 363–370 (2007)
3. Davison, B.D., Hirsh, H.: Predicting Sequences of User Actions. In: AAAI/ICML Workshop on Predicting the Future: AI Approaches to Time–Series Analysis, pp. 5–12 (1998)

4. Feng, Y., Teng, T., Tan, A.: Modeling Situation Awareness for Context-aware Decision Support. Expert Systems with Applications 36(1), 455–463 (2009)
5. Hong, J., Suh, E., Kim, S.: Context-aware Systems: A Literature Review and Classification. Expert Systems with Applications 36(4), 8509–8522 (2008)
6. Hu, M.: Model Checking for Incomplete High Dimensional Categorical Data. Ph.D. Dissertation, University of California, Los Angeles, Los Angeles, CA (1999)
7. Huýnh, T., Fritz, M., Schiele, B.: Discovery of Activity Patterns using Topic Models. In: 10th International Conference on Ubiquitous Computing, pp. 10–19 (2008)
8. Hwang, K., Cho, S.: Landmark Detection from Mobile Life Log using a Modular Bayesian Network Model. Expert Systems with Applications 36(10), 12065–12076 (2009)
9. Kim, E., Helal, S., Cook, D.: Human Activity Recognition and Pattern Discovery. IEEE Pervasive Computing 9(1), 48–53 (2010)
10. Laasonen, K., Raento, M., Toivonen, H.: Adaptive On-device Location Recognition. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 287–304. Springer, Heidelberg (2004)
11. Murphy, K.: Dynamic Bayesian Networks: Representation, Inference and Learning. Ph.D. Dissertation, University of California, Berkeley, Berkeley, CA (2002)
12. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)
13. Perl, J.: A Neural Network Approach to Movement Pattern Analysis. Human Movement Science 23(5), 605–620 (2004)
14. Petzold, J., Pietzowski, A., Bagci, F., Trumler, W., Ungerer, T.: Prediction of Indoor Movements using Bayesian Networks. In: Strang, T., Linnhoff-Popien, C. (eds.) LoCA 2005. LNCS, vol. 3479, pp. 211–222. Springer, Heidelberg (2005)
15. Rashidi, P., Cook, D., Holder, L., Schmitter-Edecombe, M.: Discovering Activities to Recognize and Track in a Smart Environment. IEEE Transactions on Knowledge and Data Engineering (2010) (in press)
16. Singla, G., Cook, D., Schmitter-Edgecombe, M.: Recognizing Independent and Joint Activities among Multiple Resident in Smart Environments. Ambient Intelligence and Humanized Computing Journal 1(1), 57–63 (2010)
17. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufman, San Francisco (2005)

Information Management for Dependability

Paul Mason

SIU International, Bangtoey, Samkhok, Thailand, 12160
paul@siu.a.th

Abstract. Dependability of a computing system is the ability to deliver service(s) that can confidently be trusted. The term embraces three parts: the attributes of, the means of attaining and threats to dependability. Owing to the severity of failure of such systems, their development is normally carefully controlled by a raft of standards and hurdles standing between the original concept and an operational system. Dependable systems are typically well understood, based around established specification techniques, fault tolerant architectures and implementation language subsets that afford dependability means. Analysis techniques that help identify dependability threats are similarly well understood. This paper considers dependability as an information management (IM) problem and proposes one possible solution.

Keywords: Dependability, Information Management, Consistency, Traceability.

1 Introduction

Nowadays, there is a growing trend towards using computers in controlling critical system services – i.e. services on which users *depend*. This phenomenon is due to several factors, including processing power, physical size, weight, and flexibility. Laprie [1] formulated a synthesis of *dependability* concepts based on its attributes (including safety, reliability and availability), its means (fault prevention, tolerance, removal and forecasting) and impairments (faults, errors and failures).

The first generation of computers (dating roughly from the late 1940's to mid-50's) were composed of unreliable components, therefore a range of techniques were used to bolster reliability, including error control codes, duplexing with comparison, and triplication with voting [2]. Work on *hardware* by [3], [4] and [5] led to what we now term fault tolerant systems. The origins of *software* fault tolerance began with work on recovery blocks [6], and later N-version programming [7].

Safety critical systems must be certified before entering service. This involves submission to the appropriate regulator of a safety case, a reasoned argument with supporting evidence that the system is acceptably safe to operate (i.e. risks associated with each hazard were either eliminated, or reduced as low as reasonably practical).

Engineers use a range of techniques during safety analysis. These divide into three categories: hazard identification (e.g. Hazard and Operability Studies) [8]; causal analysis (e.g. Fault Trees) [9]; and consequence analysis (e.g. Event Trees); again, all are established techniques. The safety case itself is generally a text-based document, although graphical notations such as [10] are beginning to find favor.

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 83–93, 2010.

Confidence in the system, and ultimately certification is at least partially dependent on confidence in consistency and traceability between results of the various analyses, and, between each set of results and the system design (system model). E.g., suppose a Fault Tree contained the Condition "cooling system failure" and that this in turn contributed to a Condition "excessive engine temperature". The Fault Tree and system model would be inconsistent if the latter did not contain a relation between the engine and cooling system entities [11]. Clearly the same levels of consistency should also apply between the development notations, and between artifacts expressed in these notations and the system model. Cleary therefore safety (and dependability generally) may be viewed as an ***information management*** (IM) problem.

Consider problems with the A400M *military transport aircraft designed by Airbus*. Note. *Designed by Airbus!* This a company whose carefully refined industrial processes made them a serious rival to Boeing in the *civil aviation* market! Problems with the A400 program arose from engine software writen by MTU Aero Engines, along with the decision by Airbus to go for civil certification, which was outside MTU's experience. Put simply, MTU contract engineers schooled in secretive defence projects, failed to meet the very different civil documentation rules required by the regulator. In correcting the problem, there was little choice but to *retrace* every step of work on the complex software. Other IM problems afflicting the project arose from stakeholder defence departments loading the A400M with a raft of extras, making it a juggernaut of conflicting requirements from the start.

In situations such as this, tool support is clearly essential. However, while some tools currently exist (e.g. [12] and [13]) we are unaware of any capable of achieving consistency at the fine granularity required to justify the level of confidence required for certification (notably confidence in consistency/traceability between the system model and analyses), nor which afford the same richness of system modeling.

A previous paper [14] presented at IAIT 2009 considered theoretical aspects of the above problem in the context of civil aviation standard ARP 4754 [15], along with a *basis* for a solution. This solution, termed *ASTrA* (*A*xiomatic approach to *S*ystem *Tr*aceability and *A*nalysis), involves an integrated tool set that emphasizes consistency by maintaining traceability links between a range of notations supporting requirements, design, implementation and testing, the results of safety analyses over artifacts expressed in these notations, and the physical system model. In this paper we expand on and demonstrate principles outlined in [14] using an in depth case study.

With this in mind, the rest of the paper is organized as follows: section two provides an overview of *ASTrA*; section three illustrates this approach using a contiguous case study from ARP 4761 [16]; section four presents our conclusions.

2 An Overview of Integrated Data Management in *ASTrA*

ASTrA addresses the lack of traceability and consistency in heterogeneous-tool environments by exporting data from the internal models of CASE tools, to a traceability *Workspace*, an interconnected set of meta-models expressed in a common modeling language. The meta-models represent selected software, systems and safety engineering notations and techniques and as such, 'receive' data exported from corresponding tools. Table 1 features (partial) meta-models capable of receiving data from Failure

Modes & Effects Analysis (FMEA) [17], Circuit Diagrams and Fault Tree Analysis tools. The models are expressed here in O-Telos and implemented using the ConceptBase Object Management System [18]. Note we concentrate in this paper on low-level representation of notations as it is their uniform format that gives us the desired traceability (and helps realize consistency with the aid of a further information model). Associations, expressed in the same common language, allow linking of meta-models (and elements), giving a seamless traceability environment.

Table 1. (Partial) FMEA, Circuit Diagram and Fault Tree Analysis Meta-models : O-Telos

FMEA Meta-model	Circuit Diagram Meta-model	Fault Tree Analysis Meta-model
Component in StructureElement, SimpleClass isA String, end ComponentType in StructureElement, SimpleClass isA String, end FailureModeRate in StructureElement, SimpleClass isA Real with constraint fmr_in_range: $ forall f/ FailureModeRate (f <= 1)$ end FailureMode in StructureElement, SimpleClass isA String end FailureEffect in StructureElement, SimpleClass isA String end ComponentFailureDescription in StructureElement, SimpleClass with has_part com_name : Component; com_type : ComponentType; com_fail_mode_description: ComponentFailureModeDescription end ComponentFailureModeDescription in StructureElement, SimpleClass with has_part com_fail_mode : FailureMode; com_fail_rate : FailureModeRate; com_fail_effect : FailureEffect has_structure com_detection : ASTrANaturalLanguageStructure; com_comment : ASTrANaturalLanguageStructure end	ConnectorPin in StructureElement, SimpleClass with has_property v : Voltage; i : Current end Connection in StructureElement, SimpleClass with has_part pin : ConnectorPin end TwoPin in … with has_property pd : Voltage; i : Current has_part p : ConnectorPin; n : ConnectorPin end ACsource in … isA TwoPin with has_property f : Frequency; va : Voltage end Resistor in … isA TwoPin with has_property r : Resistance end Capacitor in … isA TwoPin with has_property c : Capacitance end Inductor … isA TwoPin with has_property l : Inductance end CircuitDiagram in DevelopmentStructure, … has_property circuit_name : String has_element resistor : Resistor; capacitor : Capacitor; inductor : Inductor; ground : Ground; connection : Connection; connector_pin : ConnectorPin comparator : Comparator; and_gate : AndGate; dc_source : DCSource end	PreliminaryEventProfile in … isA EventProfile with has_part preliminary_budget : BudgetProbability; preliminary_rate : Rate; preliminary_exposure : Exposure; preliminary_label : EventLabel end UpdatedEventProfile in … isA EventProfile with has_part actual_probability : ActualProbability; preliminary_budget : BudgetProbability; actual_rate : Rate; actual_exposure : Exposure; actual_label : EventLabel end ActualProbability in … with has_property probability : Real has_structure annotation : ASTrANaturalLanguageStructure end BudgetProbability in … with has_property probability : Real has_structure annotation : ASTrANaturalLanguageStructure end Rate in … with has_property failure_rate : Real has_structure annotation : ASTrANaturalLanguageStructure end Exposure in … with has_property period : Real has_structure annotation : ASTrANaturalLanguageStructure end SimpleLabel in … isA EventLabel with has_part simple_description : SimpleEventDescription end

As indicated, previously the *ASTrA* framework also incorporates a system model which expresses common building blocks underlying the various notations and which maintains consistency across Workspace meta-models (see [19] for examples of how this is accomplished). Finally, the *tool2ASTrA* [19] mapping function enables data exportation from CASE tools to the Workspace meta-models. The *tool2ASTrA* interface is defined as follows:-

```
pANM tool2ASTrA (pCDS, uANM, SysMl)
```

Essentially, `tool2ASTrA` takes as its input parameters, a populated CASE tool data structure (pCDS), together with a corresponding un-populated *ASTrA* notation meta-model (uANM) and the system model (SysMl), and returns a populated *ASTrA* notation meta-model (pANM).

3 Applying *ASTrA* to an Aircraft Wheel Breaking System

This case study addresses a subset of ARP 4754/4761 assessment activities: identification and verification of safety requirements for the Brake System Control Unit (BSCU) computer for an aircraft Wheel Braking System (WBS). Requirements identification forms part of Preliminary System Safety Assessment (PSSA) of the BSCU and verification, part of System Safety Assessment (SSA) – see figure 1.

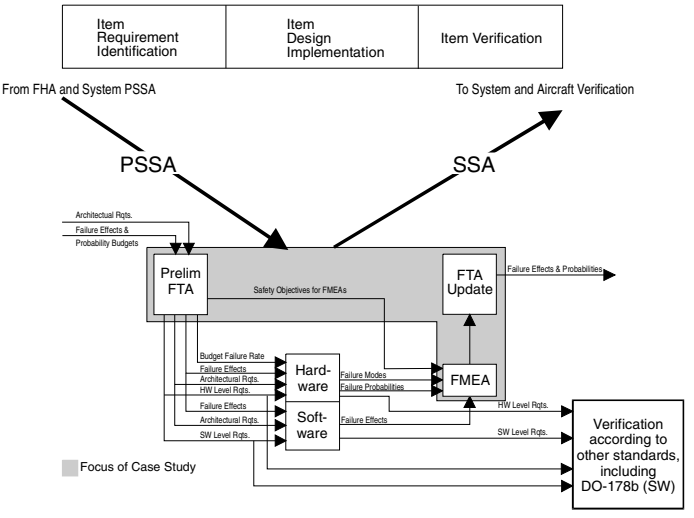


Fig. 1. ARP 4761 Assessment Process (Partial)

3.1 Overview of Wheel Braking System and Brake System Control Unit

The WBS decelerates aircraft during taxi, landing and in the event of rejected take-off. Each brake assembly has two independent sets of hydraulic pistons. One set operates from the Green hydraulic supply used in Normal Braking mode. The Alternate mode which uses the Blue hydraulic supply is on standby and is selected automatically if the Normal system fails. An Emergency braking mode is also available, although the design ensures all safety requirements can be met without regard to this system. In Normal mode, all wheels are individually braked from their own servo valves according to commands received from one of two Braking System Control Unit computers - normally BSCU1 or, if system 1 reports a failure, BSCU2.

Failure conditions for this system include ‘Inadvertent Wheel Braking’, our focus here. Such events are established by a Functional Hazard Assessment (FHA), with budget probabilities allocated according to severity. These requirements provide inputs to the (Braking System) PSSA; to determine causality, catastrophic and hazardous conditions form top events of fault trees. In turn, basic events of the system PSSA trees derive requirements that feed into an item level PSSA. This is the point where we join the example from ARP 4761.

3.2 Preliminary System Safety Assessment - Brake System Control Unit

For this case study, we assume a fault tree (not shown) developed as part of the Wheel Braking System PSSA has a basic event with the following item level requirement:-

- The probability of 'BSCU commands braking in absence of braking commands and causes inadvertent braking' shall be less that 2.5E-9 per flight.

PSSA of the BSCU subsequently serves to complete safety requirements for this component. These include quantitative requirements, one of which will now be used to demonstrate instantiation of 'preliminary' Fault Tree structure elements.

3.2.1 Background on BSCU Design

BSCU design consists of two independent systems (BSCU1 and BSCU2). Both generate necessary voltages from their respective power supplies; out of specification voltages are detected in each system by a power supply monitor (subsection 3.3.1).

Each system also contains a command and monitor channel which computes braking commands based on brake pedal inputs. Commands generated by each channel are compared, with a failure reported when a disparity is detected. Results of the power supply monitor and comparator are then provided to a system validity monitor; a failure reported by either BSCU causes that system to disable its outputs and set the system validity monitor to invalid. Each system validity monitor is provided to a BSCU validity monitor; failure of BSCU1 and BSCU2 causes a switch to the Alternate Braking System. In normal operation, BSCU1 provides the braking commands. When a failure is reported the output of BSCU2 is switched in to allow braking. If system 2 also fails, all BSCU outputs are disabled and the BSCU validity monitor set to invalid.

3.2.2 Fault Tree Analysis – Preliminary

The fault tree in figure 2 considers inadvertent wheel braking attributable to the BSCU (with requirement 2.5E-9 per flight hour) to determine feasibility of the design outlined above. In addition to safety requirements from the previous assessment phase, PSSA also takes into account operational considerations. These include average flight length (5 hours), average power-up interval (100 hours) and estimated aircraft life (100,000 hours). These provide exposure times and are used in calculating probabilities for basic fault tree events.

The fault tree in figure 2 that assumes no undetected BSCU failures can lead to inadvertent braking (BSCUUNDF); such an assumption must be proved correct through FMEA and/or Common Mode Analysis. The other branch of the fault tree, partially developed in figure 3, addresses combinations of monitored BSCU and monitor failures. Specifically, it describes analysis of the BSCU1 detectable failures causing bad data event (BSCU1DETD). In summary, this event can occur if the power supply monitor is stuck valid (BS1PSMOFV) and the power supply failure causes bad data (BSCU1PSF). An alternative path to the event is if the monitor channel always reports valid due to hardware failure (BSCU1MOFV) and command channel I/O failure causes bad data (BSCU1I/OF) or, command channel CPU hardware failure causes bad data (BSCU1CPUF).

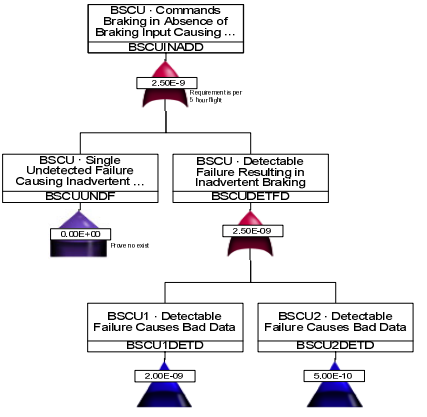


Fig. 2. BSCU Commands Braking in Absence of Brake Input - Preliminary Fault Tree

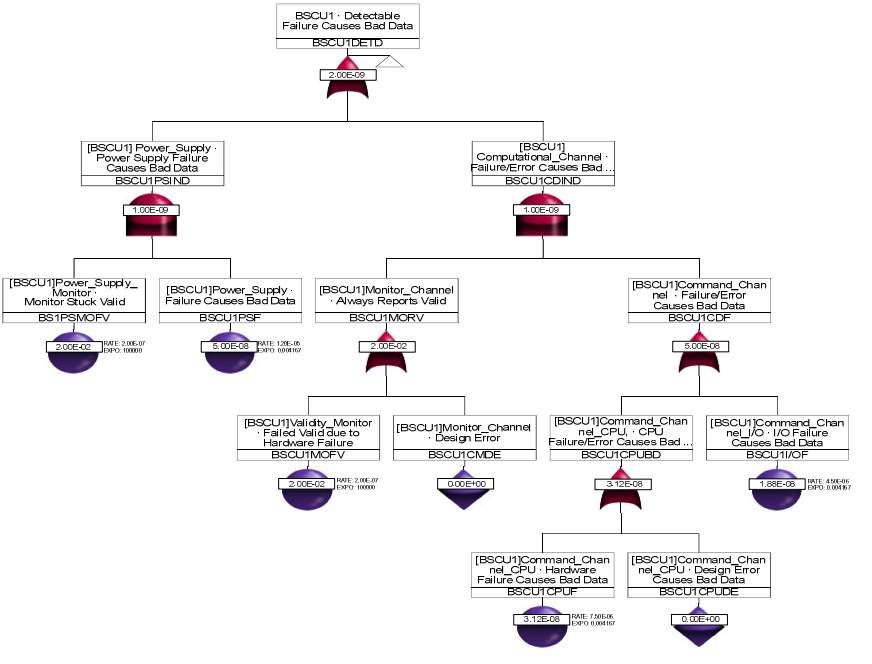


Fig. 3. BSCU Commands Braking in Absence of Brake Input– Preliminary Fault Tree (page 2)

3.2.3 Instantiation of Fault Tree Analysis Meta-model - Preliminary Fault Tree

In the following, we instantiate an O-Telos implementation of the partial Fault Tree meta-model (table 1), with the failure monitor stuck valid (BS1PSMOFV) in figure 3.

Definition of Event 'BS1PSMOFV'
EventH in Event, Token with identifier _Identifier :
"BS1PSMOFV" preliminary_profile
preliminaryProfile : EventHPreliminaryProfile end

EventHPreliminaryProfile in PreliminaryEventProfile, … with
type _Type : "bas" preliminary_budget
preliminaryBudget : EventHPreliminaryBudget
preliminary_rate preliminaryRate : EventHPreliminaryRate
preliminary_exposure preliminaryExposure :
EventHPreliminaryExposure
preliminary_label preliminaryLabel : EventHPreliminaryLabel
end

EventHPreliminaryBudget in BudgetProbability,
Token with probability _Probability : 2.00E-02 end
EventHPreliminaryRate in Rate …Token with failure_rate
failureRate : 2.00E-07 end

EventHPreliminaryExposure in Exposure, Token with
Period _Period : 100000.00 end
EventHPreliminaryLabel in SimpleLabel, Token with
simple_description simpleDescription :
EventHSimpleEventDescription end

EventHSimpleEventDescription in SimpleEventDescription
… with Entity _Entity : "Power_Supply_Monitor"
qualifiying_entity qualifyingEntity : "BSCU1"
condition _Condition : "Monitor Stuck Valid" end

3.3 System Safety Assessment - Brake System Control Unit

At this point, we assume the PSSA is sufficiently developed to allow detailed design implementation to proceed. Therefore our example resumes post-implementation by considering System Safety Assessment of the BSCU towards verification of PSSA safety objectives. This includes FMEA of the power supply and power supply monitor respectively, as well as updating of the preliminary fault tree introduced above. Artifacts produced by these activities will be used to demonstrate (partial) instantiation of the FMEA and Circuit Diagram meta-models, as well as further population of the Fault Tree Analysis meta-model.

3.3.1 Background on BSCU Power Supply Design

The BSCU1 and BSCU2 power supplies are identical in design and implementation and are located in physically remote areas of the BSCU circuit card assembly. Within each BSCU, the power supply and its monitor functions are physically independently located. Figure 4 shows a schematic of the +5 volt monitor and below it, partial population of the Circuit Diagram meta-model for the power supply monitor.

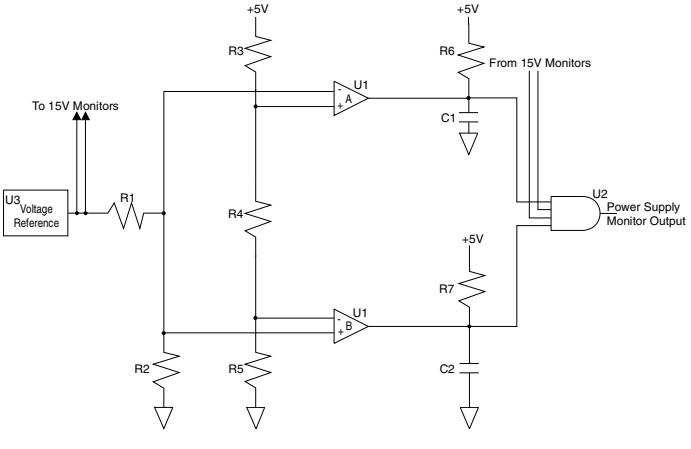


Fig. 4. BSCU + 5 Volt Power Supply Monitor Circuit Schematic

PowerSupplyMonitorCD in CircuitDiagram, Token with
-- elements from figure 4
circuit_ circuitName : "Power Supply Monitor"
resistor resistor1 : R1; resistor2 : R2; resistor3 : R3;
resistor4 : R4; resistor5 : R5; resistor6 : R6; resistor7 : R7
capacitor capacitor1 : C1; capacitor2 : C2
comparator comparator1 : U1A; comparator2 : U1B
and_gate andGate : U2
dc_source dcSource : U3 end

We return to the Circuit Diagram meta-model in section 3.4 when demonstrating traceability relations between development and assessment artifacts.

3.3.2 Failure Modes and Effects Analysis

This FMEA considers the BSCU power supply safety objectives for basic events in the 'inadvertent wheel braking' fault tree introduced earlier: Power Supply • Failure Causes Bad Data and Power Supply Monitor • Monitor Stuck Valid (we concentrate here on the latter). The FMEA is performed in two parts: i) a *functional* analysis of the entire power supply; and ii) a *piece-part* analysis of the power supply monitor which we focus on here.

3.3.2.1 Piece-Part Failure Modes and Effects Analysis. For the Basic event Monitor Stuck Valid, our scenario assumes initial analysis of the power supply monitor was part of the functional FMEA of the entire power supply. However, it was found that the total failure rate of the power supply monitor was 3.06E-7 failures per hour which does not comply with the requirement of 2.0E-7 (figure 3). Accordingly a detailed piece-part FMEA was conducted to provide better resolution into the probability of Monitor Stuck Valid. Partial results of this analysis are shown in table 2.

Several failure modes may cause a monitor to become 'stuck valid', including U1B • output open (table 2). Summed rates for contributing failures yield an actual failure rate of 1.429E-07 (including those not shown in the above), which satisfies the budget of 2.00E-07. This information will also be used in a subsequent subsection to update the BSCU preliminary fault tree.

Table 1. Piece-Part FMEA (Partial) of BSCU Power Supply Monitor

U1B	Comparator IC	output open	0.0124E-06	Monitor stuck valid	Bench test
		output grounded	0.0056E-06	Monitor trip	P/S Shut Down
		high offset voltage	0.0062E-06	Loss of monitor sensitivity	Bench test

3.3.3 Instantiation of Piece-Part Failure Modes and Effects Analysis Meta-model

We now instantiate elements from the O-Telos representation of the FMEA meta-model with information from table 2 for the failure U1B • output open.

```
Component U1B failure description

U1BFailureDescription in ComponentFailureDescription,
Token with com_name comName : "U1B"
com_type comType : "Comparator IC"
com_fail_mode_description
comFailModeDescription1 :
U1BOutputOpenFailModeDescription;
comFailModeDescription2 :
U1BOutputGroundedFailModDescription;
comFailModeDescription3 :
 U1BHighOffsetVoltageFailModeDescription end

ComponentFailureModeDescription, Token with
 com_fail_mode comFailMode : "output grounded"
com_fail_rate comFailRate: 0.0056E-6
com_fail_effect comFailEffect1 : "Monitor trip"
com_detection comDetection :
U1BOutputGroundedDetection end

U1BOutputGroundedDetection in
AstraNaturalLanguageStructure, Token
with mnls_plain_text
mnlsPlainText1 : U1BOutputGroundedDPT1 end

U1BOutputGroundedDPT1 in PlainTextNode …mnls_text
mnlsText : "Power Supply shut down." end
```

U1BOutputOpenFailModeDescription in ComponentFailureModeDescription, Token with com_fail_mode comFailMode : "output open" com_fail_rate comFailRate: 0.0124E-6 com_fail_effect comFailEffect1 : "Monitor stuck valid" com_detection comDetection : U1BOutputOpenDetection end

U1BOutputOpenDetection in AstraNatural… with mnls_plain_text mnlsPlainText1 : U1BOutputOpenDPT1 end

U1BOutputOpenDPT1 in PlainTextNode, Token with mnls_text mnlsText : "Bench test." end

U1BOutputGroundedFailModeDescription in componentFailureModeDescription, Token with com_fail_mode comFailMode : "high offset voltage" com_fail_rate comFailRate: 0.0062E-6 com_fail_effect comFailEffect1 : "Loss of monitor sensitivity" com_detection comDetection : U1BHighOffsetVoltageDetection end

U1BHighOffsetVoltageDetection in …. with mnls_plain_text mnlsPlainText1 : U1BHighOffsetVoltageDPT1 end

U1BHighOffsetVoltageDPT1in PlainTextNode, Token with mnls_text mnlsText : "Bench test." end

3.3.4 Fault Tree Analysis – Updated

The FMEA results, together with those of other techniques such as Common Mode Analysis are used to update the PSSA fault trees. This provides evidence for certification showing safety requirements for undesirable top events have been met. We now update failure rates and probabilities for the previous fault tree, with a top event addressing inadvertent braking attributable to the BSCU, and a requirement of 2.5E-09. The fragment of interest here containing event monitor stuck valid whose O-Telos code is show as follows.

EventH in Event, Token with updated_profile updatedProfile : EventHUpdatedProfile end

EventHUpdatedProfile in UpdatedEventProfile, Token with type _Type : "bas" actual_probability actualProbability : EventHActualProbability preliminary_budget preliminaryBudget : EventHPreliminaryBudget actual_rate actualRate : EventHActualRate actual_exposure actualExposure : EventHPreliminaryExposure actual_label actualLabel : EventHPreliminaryLabel end

EventHActualProbability in ActualProbability, Token with probability _Prob : 1.43E-2 end

EventHActualRate in Rate, Token with failure_rate failRate : 1.43E-7 end

Failure effects and probabilities from the item level SSA are subsequently used to verify system level analyses (including fault trees), and thence aircraft level analyses.

3.4 Trace Relations

Trace relations may be added to Workspace meta-models to allow navigation between these models and their elements. Two in particular are summarised in table 3.

Table 3. Summary of Traceability Relations (from Figure 6)

SOURCE		RELATION TYPE	TARGET	
Model	**Element**		**Model**	**Element**
Power Supply Monitor Circuit Diagram	U1B (Capacitor)	*Assessed-By*	Failure Modes & Effects Analysis Table	U1B Failure Description (Component, Type and Failure Modes)
Failure Modes & Effects Analysis Table	output open (Failure Mode Description)	*Contributes-to*	Fault Tree Analysis	Power Supply Monitor • Monitor Stuck Valid (Event Profile)

3.4.1 Instantiation of Trace Relations

To demonstrate application of these relations, we instantiate the appropriate O-Telos classes originally introduced in [19]. Our first example is an AssessedBy relation between U1B, a comparator element of the Power Supply Monitor Circuit

Diagram - and the corresponding description of failures for that particular component from the piece-part FMEA table (U1BFailureDescription).

AssessedBy01 in AssessedBy, Token with from_entity fromEntity : U1B to_entity toEntity : U1BFailureDescription end

Our second example populates a ContributesTo association between a failure effect (U1BOutputOpenFailModeDescription from the piece-part FMEA) and a fault tree updated profile (EventHUpdatedProfile).

ContributesTo01 in ContributesTo, Token with from_entity fromEntity : U1BOutputOpenFailModeDescription to_entity toEntity : EventHUpdatedProfile end

4 Conclusions

This paper has demonstrated application of our ASTrA traceability framework. Specifically, we illustrated means of realizing traceability between meta-models for Circuit Diagram, Fault Tree and FMEA (included in part here and shown fully in refs [14, 19]) using extracts of a case study from an aerospace industry standard. Potentially, data expressed using these notations and techniques may reside in different CASE tools, making it difficult to maintain traceability and consistency between them. By exporting the data to meta-models expressed in a uniform format (i.e., in a common language), links (expressed in the same language) can be inserted that capture traceability dependencies.

The novelty of ASTrA lies in two areas:-

1. A framework for traceability and consistency for safety engineering projects.
2. Meta-models for a representative set of safety assessment techniques, as well as development and project management notations and techniques used in safety engineering.

Finally, we highlight related work, notably [20] which provides tool support for managing traceability and consistency of safety related information. This work is closest in spirit to ASTrA, however it employs its own custom editors for FTA, FMEA, system modeling, etc. While this provides one extremely valid solution to the problems of traceability and consistency, practitioners may not be willing to relinquish their existing set of CASE tools for bespoke alternatives.

References

1. Laprie, J.C.: Dependable Computing and Fault Tolerance: Concepts and Terminology. In: Proc. 15th IEEE Int. Symp. on Fault-Tolerant Computing (1985)
2. Reliability and Requirements. In: Procs. Eastern Joint Computer Conference (December 1953)
3. von Neumann, J.: Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components. Annals of Math Studies 34 (1956)
4. Moore, E.F., Shannon, C.E.: Reliable Circuits Using Less Reliable Relays. J. Franklin Institute 262, 191–208, 281–297 (1956)
5. Pierce, W.H.: Failure-Tolerant Computer Design. Academic Press, London (1965)

6. Randell, B.: System Structure for Software Fault Tolerance. IEEE Trans. on Software Engineering, SE 1, 1220–1232 (1975)
7. Avizienis, A., Chen, L.: On the Implementation of N-version Programming for software Fault Tolerance. In: Proc. IEEE COMPSAC 1977, pp. 149–155 (November 1977)
8. Chemical Industries Association, A Guide to Hazard and Operability Studies (1977)
9. Vesely, W.E.: Fault Tree Handbook, NUREG–0492 (1992)
10. Kelly, T., Weaver, R.: The Goal Structuring Notation, A Safety Argument Notation. In: Proc. of Dependable Systems and Networks (2004)
11. Liu, S., McDermid, J.A.: A Model-Oriented Approach to Safety Analysis Using Fault Trees and a Support System. Journal of Systems and Software 35(2), 151–164 (1996)
12. Knudsen, J., Smith, C.: Common Cause Modeling in SAPHIRE. In: Procs. 17th International System Safety Conference, Florida, August 16-21 (1999)
13. Relex Reliability Studio, Demonstration Version (2009), http://www.relex.com/
14. Mason, P.: An Axiomatic and Object-based Approach to Tracing Safety Properties in the Context of ARP 4754. In: Papasratorn, B., et al. (eds.) IAIT 2009. CCIS, vol. 55, pp. 81–95. Springer, Heidelberg (2009)
15. Certification Considerations For Highly-Integrated or Complex Aircraft Systems, ARP 4754 (1996)
16. Conducting the Safety Assessment Process on Civil Airborne Systems and Equipment, ARP 4761 (1996)
17. Bussolini, J.: High Reliability Design Techniques applied to the Lunar Module, London. Lecture Series Avionics Systems, vol. 47 (September 1971)
18. Jarke, M., Gallersdorfer, R., Jeusfeld, M., Staudt, M., Eherer, S.: ConceptBase: A Deductive Object Base for Meta Data. Journal of Intelligent Info. Sys., Mar., 167–192 (1995)
19. Mason, P.: Managing Complexity in ICT Systems Development. Int. Journal of Information Technology and Management 7(3), 264–282 (2008)
20. Wilson, S., et al.: Safety Case Development: Current Practice, Future Prospects. In: Proc. 1st ENCRESS Conf., Bruges, Belgium (1995)

A Prototype for the Support of Integrated Software Process Development and Improvement

Nalinpat Porrawatpreyakorn[1], Gerald Quirchmayr[1,2], and Wichian Chutimaskul[3]

[1] University of Vienna,
Faculty of Computer Science, Department of Distributed and Multimedia Systems,
Liebiggasse 4, A – 1010 Vienna, Austria
a0848231@unet.univie.ac.at, Gerald.Quirchmayr@univie.ac.at
[2] University of South Australia, School of Computer and Information Science,
Mawson Lakes, SA 5095, Australia
Gerald.Quirchmayr@unisa.edu.au
[3] King Mongkut's University of Technology Thonburi, School of Information Technology,
126 Prachauthit Rd., Tungkru, Bangkok, 10140, Thailand
wichian@sit.kmutt.ac.th

Abstract. An efficient software development process is one of key success factors for quality software. Not only can the appropriate establishment but also the continuous improvement of integrated project management and of the software development process result in efficiency. This paper hence proposes a software process maintenance framework which consists of two core components: an integrated PMBOK-Scrum model describing how to establish a comprehensive set of project management and software engineering processes and a software development maturity model advocating software process improvement. Besides, a prototype tool to support the framework is introduced.

Keywords: Software Development, Project Management, Software Process Improvement, Software Process Maintenance.

1 Introduction

A software development process (hereafter referred to as a software process) is the engineering and management framework consisting of a set of phases, stages, methods, activities, and tools used to develop and maintain software [1]. It is viewed as a vehicle to deliver the quality of software [2]. The software process should thus be efficient. The efficiency requires not only the establishment but also the continuous improvement of integrated project management and of the software process [3, 4]. Currently agile methods are ubiquitously applied for the rapid delivery of quality software; however, they are not efficient enough in the managerial sense. Issues such as limited support for outsourcing, developing with large teams, developing software that demands high quality control, and distributed development environments [5, 6, 7, 8] often remain uncovered. As this research aims at minimizing changes of the processes which software development teams are already familiar with, the Project Management Body of Knowledge (PMBOK) and Scrum are accordingly used as basis for

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 94–105, 2010.

an integrated PMBOK-Scrum model. PMBOK is the broadest and most widely used standard reference of industry best practices for project management [9]. Scrum has emerged as the most successful agile development process for organizations and software development teams [10]. Moreover, Capability Maturity Model Integration (CMMI) as a Software Process Improvement (SPI) de facto method [11] has proved that the effort put into this method can assist in producing high quality software, reducing cost and time, and increasing productivity [12, 13, 14, 15]. Many studies also emphasize the importance and the use of critical success factors (CSFs) in SPI rather than CMMI key process areas (e.g., CSFs help to extend the boundaries and increase the value of SPI) [15, 16, 17, 18] and have confirmed its value [15, 17]. Hence, CMMI and CSF are used as basis for a Software Development Maturity (SDM) model. In this paper, both models are core components of a software process maintenance framework, which in this context means a framework for software process establishment and improvement.

Many organizations face either unfulfilled promises about software quality gained from applying software engineering approaches, or the inability to manage the software process realized as their fundamental problem [19]. The search for solutions to this barrier has continued for decades. Consequently, the answer to "How to improve software process development?" should help software development teams by acting as guidance for quality software development. So far, the identified CSFs and the proposal of the framework relating to this question have been presented in [20]. In this contribution, the framework with its core components and a prototype tool are presented. Practical tests of our approach with industry partners are in an early planning stage.

2 A Software Process Maintenance Framework

In order to consistently deliver quality results, an efficient software process requires both, the establishment and the continuous improvement of an integrated software management and engineering process. This paper thus proposes a software process maintenance framework as depicted in Fig. 1.

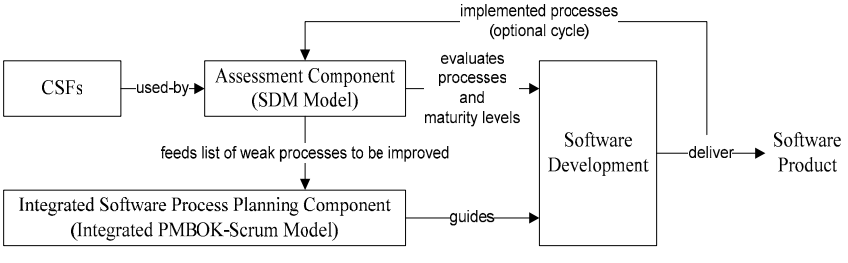


Fig. 1. The suggested software process maintenance framework

The framework has adequately paid attention to the "what", "how", and "how good" to implement the software process through an SDM model and an integrated PMBOK-Scrum model. The SDM model as an assessment component is created with a threefold objective: to appraise an organization's current software process through

the identified CSFs [20], to get the current maturity level rating from the model, and to identify which processes demand immediate and sustainable improvement. The SDM model is based on a CMMI staged representation and CSF approaches and adapted from [15] as illustrated in Fig. 2. In the model there are three dimensions: maturity stage, CSFs, and assessment. For the maturity stage dimension, we have at this stage adopted four CMMI maturity levels: initial, managed, defined, and optimizing. The main reason not to replicate the CMMI maturity level-4 "quantitatively managed" is that the two key CMMI practices of establishment and maintenance of quantitative objectives for the process, and stabilization of the performance of one or more sub-processes to determine its ability to achieve are not compatible with agile best practices [21]. Besides, there is no CSF cited in the literature that directly relates to this level [15].

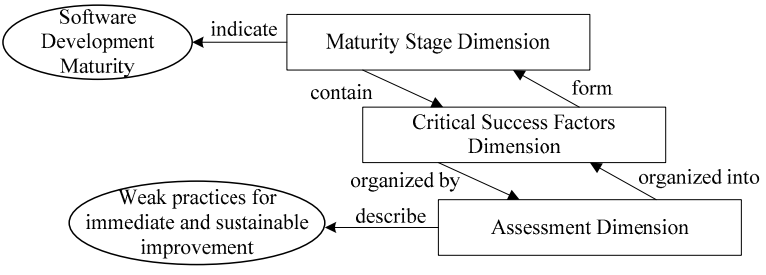


Fig. 2. An SDM model structure (adapted from [15])

At the starting point of this research, the goal was to identify CSFs of software development (see [20]), which are directly used in the CSFs dimension. We have analyzed 54 international sources, the most important ones being [5, 6, 14, 15, 16] (all references to these sources are available from the corresponding author) for our CSF categorization and the design of lists of practices. CMMI process areas are organized into categories across the five maturity levels [22]. We have hence adopted this mechanism and categorized the identified CSFs into three categories: foundation, standardization, and support. There is no category for the level-1 "initial" since this level does not have to be achieved due to its chaotic characteristic (similarly to CMMI). At the maturity level-2 "managed", basic project management processes, necessary process discipline, and commitments among key stakeholders are established. Thus, the foundation category containing CSFs that are the foundation for all subsequent levels can be linked to this level. These CSFs are management commitment, project management process, project type, user involvement, and training support. At the maturity level-3 "defined", the software management and engineering processes are standardized and integrated into a standard software process. Accordingly, the standardization category containing CSFs that support the design of systematic structures can be linked to this level. These CSFs are organizational environment, team environment, agile software engineering process, appropriate methods and tools, team capability, team size, and data quality. Moreover, continuous process improvement must be enabled; therefore, the support category, which contains the reviews factor to support continuous SPI activities, can be linked to the maturity

level-4 “optimizing”. In order to guide how to assess and implement CSFs, a list of practices scrutinized and/or recommended in the literature above has been designed for each CSF. Table 1 shows an example of lists of CSFs’ practices.

Table 1. Lists of CSFs’ practices

CSF	List of Practices
Agile software engineering process	1. A project has been established with well-defined coding standards up front 2. A project has been established by pursuing simple design 3. A project has been established with rigorous refactoring activities 4. A project has been established with right amount of documentation 5. A project has been established with correct integration testing 6. A project has been established with short increments 7. Most important features have been first delivered 8. Cost evaluation has been done up front 9. Risk analysis has been done up front
Organizational environment	1. Cooperative organizational culture has been established instead of hierarchal culture 2. Oral culture placing high value on face-on-face communication has been established 3. Agile methodology has universally been accepted in the organization 4. All the key stakeholders are involved in development and improvement initiatives 5. Management has provided strong leadership-collaboration 6. Facility with proper agile-style work environment has been established 7. Reward system appropriate for agile software development has been promoted among the management and team members
Team environment	1. Collocation of the whole team has been established 2. Coherent and self-organizing teamwork has been established 3. A project has been established with no multiple independent teams 4. A process has been established to monitor the progress of each team 5. A process has been established to collect and analyze the feedback data from each team and to extract the main lessons learned 6. A process has been established to distribute the lessons learned to the relevant stakeholders and team members 7. Team members are aware of their roles and responsibilities during software development and improvement

Last, in the assessment dimension we have adapted an assessment instrument successfully developed and tested at Motorola (see [23]) to measure SDM since it is a normative instrument designed to be adapted [24]. The instrument has three evaluation dimensions: *Approach*, key criteria here are the organization commitment to and management support for the practice, and the organization’s ability to implement the practice; *Deployment*, key criteria here are the breadth and consistency of practice implementation across project areas; and *Results*, key criteria here are the breadth and consistency of positive results across project areas. Sets of key criteria for scores of 0 to 10 are guided for rating. To calculate, under each CSF each practice is weighted by three-dimensional scores in integer between 0 and 10. The three-dimensional scores of each practice are added, divided by 3, and rounded up. All obtained practice scores

are then rolled into an average score for each CSF. Any CSF with an average score falling below the threshold is deemed a weakness. The threshold is initially set to 7 as guided by [23]; however, it can be reset to fit an organization's current circumstance. To achieve a certain maturity level, all CSFs belonging to that maturity level should have an average score of the threshold or higher. Table 2 shows an example of a CSF's evaluation.

Table 2. A CSF evaluation (average score = 6+8+7+7/no. of practices = 28/4 = 7)

Management Commitment	Three-Dimensional Scores			Average Score
	Approach	Deployment	Results	
1.Management provides strong commitment and presence	6	6	6	6
2.Management supports the software development	8	7	8	8
3.Management is willing to participate in assessment and development activities	7	7	6	7
4.Management is committed to provide training and resources	8	7	6	7
Average Score for Management Commitment				7

For the second component, an integrated PMBOK-Scrum model as an integrated software process planning component is constructed to assist in establishing, designing, and planning an integrated software process. The model (as illustrated in Fig. 3 using Unified Modeling Language notation) is composed by three layers (i.e., a managerial layer and a production layer which are derived from the distinction of concepts of PMBOK and Scrum respectively, and an integration layer which is derived from an overlapping of concepts). Besides, a relationship of concepts is created as an association between related classes. The model is developed based on the same kind of approach in that it is made for the integration of, e.g., PMBOK and RUP (Rational Unified Process) [25, 26] and PMBOK and OPEN (Object-oriented Process, Environment and Notation) [26], and explained similarly to [25]. We begin with the overlapping concepts. In the integration layer, both PMBOK and Scrum have a **project** concept which has a collection of **phases**. Each phase has a set of **activities**. Given activities are defined as managerial or productive entities. They can be decomposed into **tasks** and can have **dependencies** among them, which allow the definition of the order in which they can occur inside the project. The activities can also be supported by **guidance** which includes **tools** and **techniques**. Although Scrum has a project management perspective, in the integrated model its project management concepts are moved into the managerial layer. **A managerial activity** typically produces **work products**, i.e., **deliverables**. In contrast, a **productive activity** produces **artifacts** and incremental products.

Each work product has its **type** and **version**. Like activities, **roles** are categorized into **managerial roles** and **productive roles** to perform managerial and productive activities, respectively. Any given activity has a single role responsible for it as argued by good project management rules [25, 27]. To each association between a role

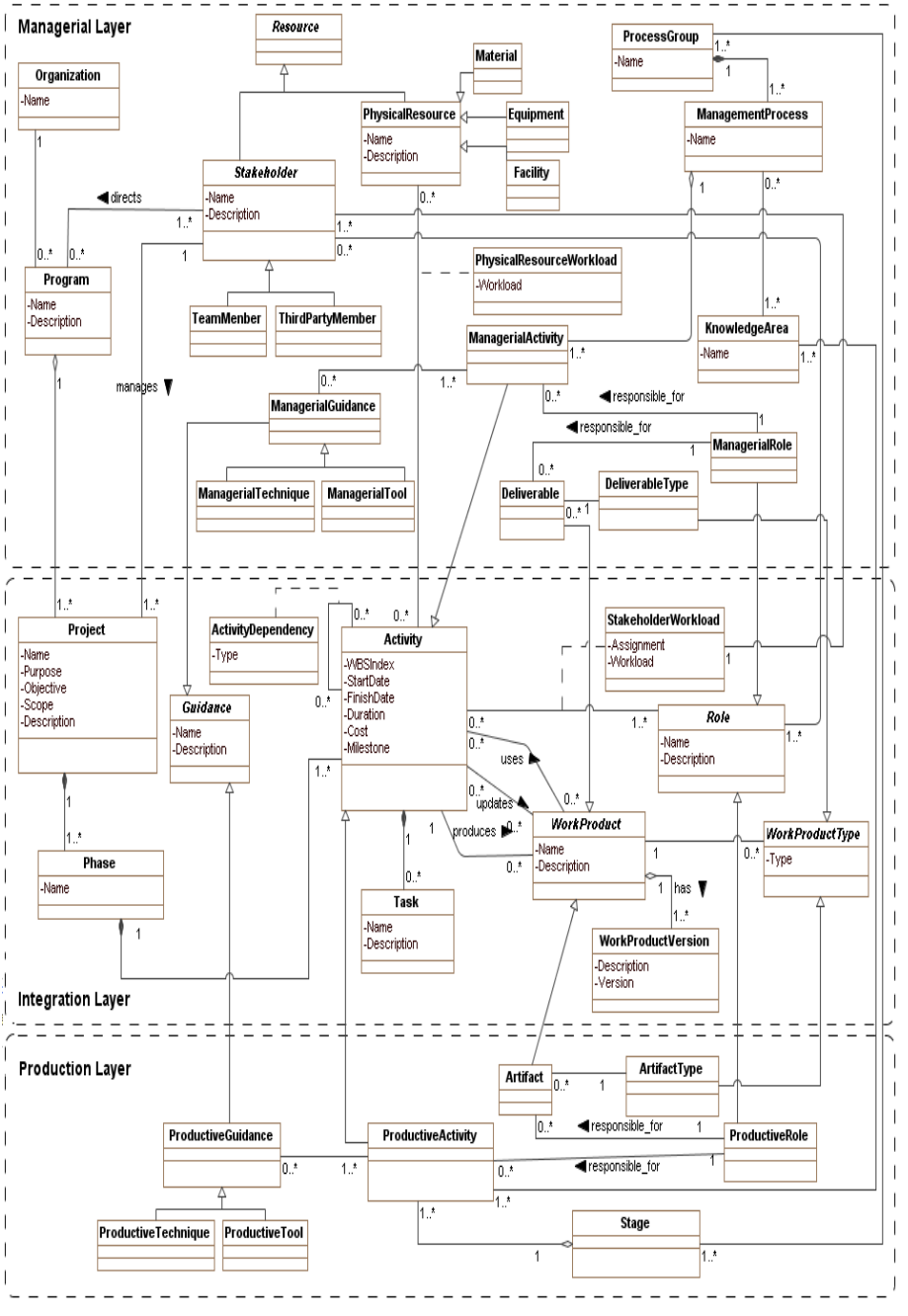


Fig. 3. An integrated PMBOK-Scrum model

and its activities a stakeholder must be present, including the stakeholder's workload which is an attribute in the **StakeholderWorkload** class in that relation. Similarly, physical resources and their workload which is an attribute in the **PhysicalResourceWorkload** class are present. An activity may produce, update, or use a work product which is presented as **produces**, **updates**, or **uses** associations, respectively. In the productive layer, a productive role is responsible for **productive activities** which associate one or more management **knowledge areas**. Each productive activity is organized into one **stage,** which associates one or more management process groups. In the managerial layer, an **organization** has a collection of projects organized into a **program**. Each program is run by appropriate resources. The **Resource** class consists of **Stakeholder** and **PhysicalResource**. Stakeholders can comprise both inside and outside company, called **team members** and **third party members**, respectively. Physical resources include **materials**, **equipment**, and other **facilities** that can be utilized to support managerial and productive activities in the project. Besides, each **managerial process** consists of one or more managerial activities and is organized into one **process group** and one knowledge area. Furthermore, to assure the model's appropriateness and consistency, the following set of constraints is thus formalized through the Object Constraint Language: (a) a program must have a director; a stakeholder who is a program director must have a managerial role; (b) a project must have only one project manager; a stakeholder who is a project manager must have a managerial role; (c) an activity flow must not result in a cycle (for example, activity A is a prerequisite for activity B and activity B is also a prerequisite for activity A); (d) the same activity can only either produce, update, or use the same work product; they must be performed in different activities; (e) each activity must be performed by at least one role and have only one stakeholder responsible for it; the stakeholder must also be compatible with that role; (f) a managerial activity can not produce or update a productive work product, except only a managerial work product; however, this activity can use a productive work product; (g) a productive activity can not produce or update a managerial work product, except only a productive work product; however, this activity can use a managerial work product; and (h) an activity can update or use a work product only in case it has already been created by a predecessor activity; otherwise it first needs to produce that work product. To facilitate the use of framework, the following section presents a prototype tool.

3 A Software Process Assessment and Development Tool

A technique is a defined systematic procedure employed to perform an activity to produce a product, and may employ associated software development tools which are programs or applications [9]. The framework providing an integrated software process development and improvement technique might be too complex without the right tools. A set of associated tools is hence needed to ensure the quality of software processes and products. Under the framework foundation, we designed a prototype tool called SPAD (Software Process Assessment and Development) to assist in the assessment, improvement and establishment of a well-defined integrated software management and engineering process. The prototype is being developed as a Web-based application, using Java language and a MySQL database. It should emphasize that

SPAD can serve several software development projects with independent development teams or organizations working at the same time. All the information pertinent to those projects must be managed by only their owners. When logging on our SPAD tool, SPAD hence first authenticates a user, and then guides how to appraise, improve and establish an integrated software management and engineering process in five steps. Fig. 4 illustrates the SPAD functionality using a Use-case diagram.

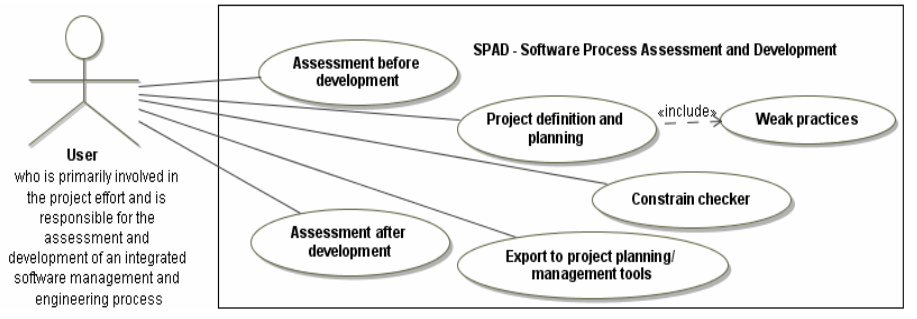


Fig. 4. A Use-case diagram showing the main SPAD functionality

Through a Use-case diagram, a user (e.g., a project manager) who is primarily involved in the project effort can gain considerable insight into the organization's current maturity by evaluating it in the project management and software engineering environments as a prerequisite to defining and planning the process required. To do so, in the assessment module, the user can first assess the identified CSFs through their lists of practices required by the SDM model. SPAD then calculates with the aforementioned calculation logic, and summarizes the obtained maturity level and scores in scoring worksheets, the overall status of the CSFs in bar charts, and weak practices in tables. The obtained results should assist the user in understanding the organization's current situation, e.g. by providing help with identifying strong practices and weak practices. For supporting the user in the planning of a project, weak practices will automatically be fed into the project definition and planning module as required practices for improvement.

Second, in the project definition and planning module the user can define the information required by the integrated PMBOK-Scrum model. The entire information consists of organization, program, project, phase, stage, work product, role, activity, guidance, stakeholder, physical resource, managerial knowledge area, managerial process group, managerial process, working time, and work breakdown structure code information. This information and the weak practices from the assessment module are used to plan the project. The module is at this stage designed to assist the user in developing plans, assigning resources to tasks, and analyzing workloads. Fig. 5 shows a screenshot of the current version of this module.

Third, the defined process is validated by the constraint checker module to assure appropriateness and consistency. The module contains the aforementioned restrictions proposed by the integrated PMBOK-Scrum model. The validation tree-based results will be shown in order to easily track an inappropriate process.

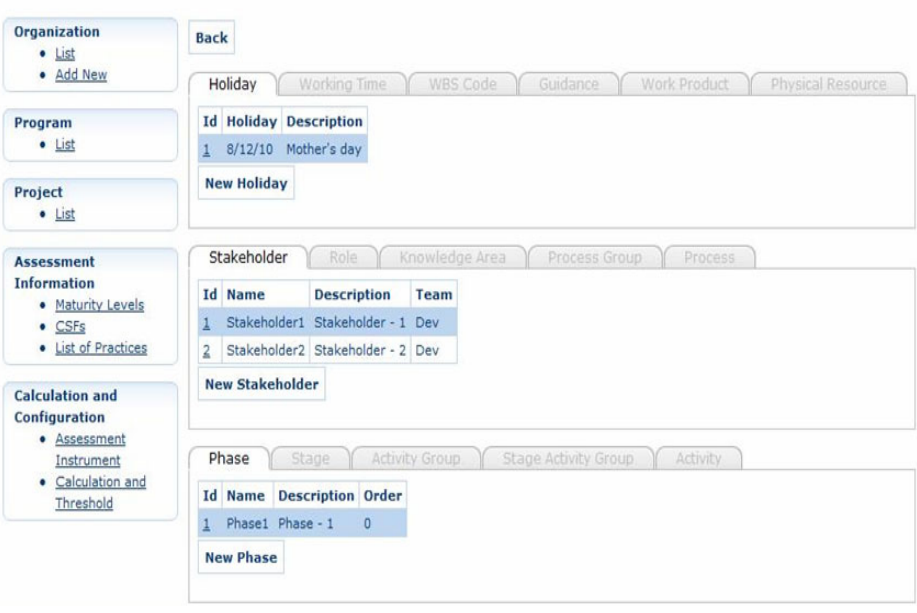


Fig. 5. A screenshot of project definition and planning module

Fourth, in the export module, the validated project is prepared in a form of a standardized MS Project 2003 XML file format for export to the organization's project planning tools. Owing to the ubiquitous XML standard, the project XML file is compatible with several commercial tools, e.g., MS Project and Oracle Primavera, and freeware tools, e.g., OpenProj, GNOME Planner, Gantter, and Basecamp. The main reason for the need of other project planning/management tools is that SPAD provides the limited functionality as aforementioned. For instance, in case the needs of abilities to track progress, manage budgets, or visualize schedules in Gantt charts, the user should export the project to their suitable tools.

Fifth, the user can again perform the implemented process appraisal in the assessment module to compare the performance of the before- and after- software process development. As stated by [28], the process improvement cycle involves three stages: process measurement, attributes of the current process are measured; process analysis, the current process is assessed and weaknesses are identified; process change, changes to the process are introduced. After software development, these three stages should be gone through. One of the criteria that can be used for evaluating the performance of the process improvement is the obtained higher scores of CSFs or maturity levels. This module should therefore assist the user in considering the performance of the process improvement. Fig. 6 shows a screenshot of the before- and after-implemented CSFs comparison.

This is the first prototype and will therefore need further evaluation and improvement. Consequently, our approach and tool will be tested through case studies in the telecommunications industry in Thailand in the last quarter of 2010.

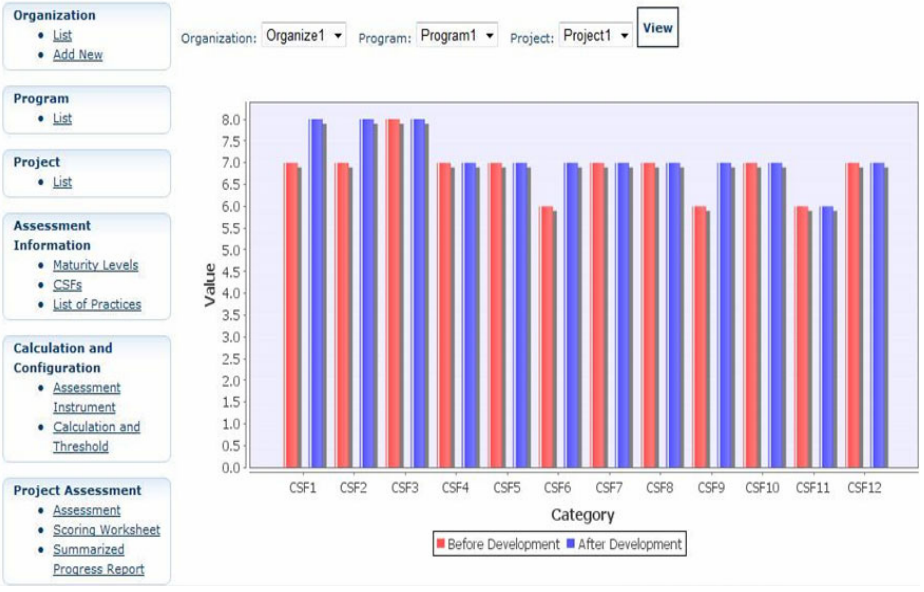


Fig. 6. A screenshot of the before- and after- implemented CSFs comparison

4 Conclusion and Future Work

One of the factors that play a central role in quality software development is an efficient software process. In this contribution, we have presented a software process maintenance framework which consists of two components: an SDM model aiming at guiding which processes demand immediate actions and an integrated PMBOK-Scrum model aiming at describing how to implement a comprehensive set of software management and engineering processes. Besides, we have created a prototype tool to support the use of framework. However, the framework needs further evaluation and improvement. Usability is defined as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [29]. Hence, a usability analysis is required. To measure the usability of the framework, two criteria (i.e., ease of use and usefulness which are profoundly rooted in attitude towards use [30]) will be used. Moreover, the case study methodology is well suited for software engineering research due to contemporary phenomena in its natural context [31]. Practical tests of our approach will be carried out through empirical studies in cooperation with companies in the telecommunications industry in Thailand in the last quarter of 2010. The case studies will be performed with a threefold objective: to test the validity of our approach; to highlight deficient areas of our approach; and last to illustrate the practicality of our approach. Data collection will also be done through interviews and onsite observations.

Acknowledgments. The authors would like to thank ÖAD, the Austrian Agency for International Cooperation in Education and Research, and the Higher Education

Commission of Thailand for supporting this work in the form of a scholarship for Nalinpat Porrawatpreyakorn.

References

1. Ambler, S.W.: The Unified Process Elaboration Phase: Base Practices in Implementing the UP. CMP, Kansas (2000)
2. Lehman, M.M.: Why is process important? In: ICSP-1, p. 4 (1991)
3. Abrahamsson, P., Warsta, J., Siponen, M.T., Ronkainen, J.: New Directions on Agile Methods: A Comparative Analysis. In: ICSE-25, pp. 224–254 (2003)
4. Pualk, M.C., Curtis, B., Chrissis, M.B., Weber, C.V.: Capability Maturity Model^{SM} for Software, Version 1.1. Software Engineering Institute Technical Report (1993)
5. Highsmith, J., Cockburn, A.: Agile Software Development: The Business of Innovation. IEEE Computer 34, 120–122 (2001)
6. Ionel, N.: Critical Analysis of the Scrum Project Management Methodology. In: 4th International Economic Conference on European Integration - New Challenges for the Romanian Economy Oradea, Romania, pp. 435–441 (2008)
7. Shalloway, A., Beaver, G., Trott, J.R.: Lean-Agile Software Development: Achieving Enterprise Agility. Addison-Wesley Professional, Reading (2009)
8. Turk, D., France, R., Rumpe, B.: Limitations of Agile Software Processes. In: Wells, D., Williams, L. (eds.) XP 2002. LNCS, vol. 2418, pp. 43–46. Springer, Heidelberg (2002)
9. PMI: A Guide to the Project Management Body of Knowledge (PMBOK Guide), 4th edn. Project Management Institute, Inc., Pennsylvania (2008)
10. Danube Technologies Sees Strong Growth/Scrum Emerges as Leading Method for Agile Software Development, `http://www.agilejournal.com`
11. Solingen, R.: Measuring the ROI of Software Process Improvement. IEEE Software 21, 32–38 (2004)
12. Yamamura, G.: Software Improvement Satisfied Employees. IEEE Software, 83–85 (1999)
13. Pitterman, B.: Telcordia Technologies: The Journey to High Maturity. IEEE Software, 89–96 (2000)
14. Jiang, J.J., Klein, G., Hwang, H., Huang, J., Hung, S.: An Exploration of the Relationship between Software Development Process Maturity and Project Performance. Information & Management 41, 279–288 (2004)
15. Niazi, M., Wilson, D., Zowghi, D.: A Maturity Model for the Implementation of Software Process Improvement: An Empirical Study. Systems and Software 74, 155–172 (2005)
16. Fitzgerald, B., O'Kane, T.: A Longitudinal Study of Software Process Improvement. IEEE Software, 37–45 (May/June 1999)
17. Somers, T., Nelson, K.: The Impact of Critical Success Factors Across the Stages of Enterprise Resource Planning Implementations. In: HICSS 34, p. 8016 (2001)
18. Khandelwal, V., Natarajan, R.: Quality IT Management in Australia: Critical Success Factors for 2002. Technical Report No. CIT/1/2002, University of Western Sydney (2002)
19. Defense Science Board Washington DC: Report of the Defense Science Board Task Force on Military Software (1987)
20. Porrawatpreyakorn, N., Quirchmayr, G., Chutimaskul, W.: Requirements for a Software Process Maintenance Framework for Executive Information Systems in the Telecommunications Industry. JGMR 6, 7–17 (2010)

21. Jain, A.: Annual Research Review & Executive Workshop Post Workshop Progress Report (2002)
22. SEI: Capability Maturity Model Integration for Software Engineering (CMMI-SM), Version 1.1 (2002)
23. Daskalantonakis, M.K.: Achieving Higher SEI Levels. IEEE Software 11(4), 17–24 (1994)
24. Niazi, M.: An Instrument for Measuring the Maturity of Requirements Engineering Process. In: Bomarius, F., Komi-Sirviö, S. (eds.) PROFES 2005. LNCS, vol. 3547, pp. 574–585. Springer, Heidelberg (2005)
25. Callegari, D.A., Bastos, R.M.: Project Management and Software Development Processes: Integrating RUP and PMBOK. In: ICSEM 2007 (2007)
26. Rosito, M.C., Callegari, D.A., Bastos, R.M.: Gerência de Projetos e Processos de Desenvolvimento de Software: uma proposta de integração, http://www.sbc.org.br/sbsi/2008/doc/SBSI2008ArtigoGerenciaProjetos.pdf
27. Schwalbe, K.: Information Technology Project Management, 2nd edn. Thomson Learning, Canada (2002)
28. Sommerville, I.: Software Engineering, 7th edn. Pearson Addison Wesley, London (2004)
29. ISO: ISO DIS 9241-11: Guidance on Usability (1994)
30. Davis, F.D.: Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly 13, 319–340 (1989)
31. Runeson, P., Höst, M.: Guidelines for Conducting and Reporting Case Study Research in Software Engineering. Empir. Software Eng. 14, 131–164 (2009)

The Effects of Organizational Experiences on Career Satisfaction of IT Postsecondary Teachers in Thailand

Theerapath Prawatrungruang, Pruthikrai Mahatanankoon, James Wolf, and Joaquin Vila-Ruiz

Illinois State University, School of Information Technology,
Campus Box 5150, Normal, Illinois, USA
{tpawat,pmahata,jrwolf,javila}@ilstu.edu

Abstract. The quality of information technology education depends heavily on the teaching commitment of Thai IT educators. With many universities competing for academic excellence, the quality of teaching and career satisfaction among Thai educators is often overlooked. To understand how different organizational experiences influence career satisfaction and teaching commitment, data collected from Thai IT educators reveals that organizational acceptance, training programs, and academic support influence career satisfaction. Thai universities can focus on strategies that will enhance these organizational experiences, which can lead to higher quality IT educators in Thailand.

Keywords: career satisfaction, organizational experiences, post-secondary education, Thailand, teaching commitment.

1 Introduction

Information Technology (IT) is an important part of modern industries and organizations. Nearly every organization relies on IT to efficiently accomplish its tasks. Thus, there is a strong demand for IT professionals and educators in Thailand, but higher salary and corporate market demands for new IT graduates often lead to a shortage of IT educators especially among Thai private universities. According to the Office of the Higher Education Commission (www.inter.mua.go.th) and its Ministry of Education, Thailand is experiencing teacher shortages in all levels from elementary school to higher education. This problem generally stems from the lack of quality teacher education and research support.

To address this concern, Thailand's National University Project seeks to strengthen its educational system through limited financial incentives—an effort to promote teaching commitment and quality research. In 2008, Thammasat University reported that the percentage of educators with doctoral degree is less than fifty percent and only one in fifty has the rank of full professor. Considering a country with roughly an area of 198,000 square miles (nearly equivalent to the size of France, or the area of the states of Montana and Arkansas combined), Thailand has at least 166 higher education institutions under the supervision of the Commission on Higher Education, Ministry of Higher Education (www.moe.go.th). With a growing number of universities and students, the country needs to retain and produce more quality teachers.

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 106–115, 2010.

Moreover, as these universities try to attract students by improving their reputation and academic quality, they often overlook the impact of organizational experiences and career satisfaction among Thai educators.

The higher education job market in Thailand is quite different than that of the U.S., however. According to Computing Research Association (www.cra.org), U.S. professors change jobs due to the following reasons: access to quality graduate students, departmental morale/culture, departmental ranking/reputation and salary. Other considerations of becoming an educator in the U.S. are attaining tenure, availability of sabbatical leave, research funding, and the balance between teaching, research and service.

In Thailand, socio-cultural influences often play an important role in how one chooses a profession. The Thai word for "teacher" is *Kru* (although *Ajarn*—a person who gives knowledge—is typically being used in higher education), which derive its meaning from "industrious". *Kru* or *Ajarn* inevitably receive social respect and altruism. But social benevolence also bestows *Kru* and *Ajarn* with surmountable challenges, including shifting teaching hours, lesser pay and other job-related responsibilities—research and public services. This is apparently why the best and brightest Thai IT graduates generally prefer working for multi-national or private companies, who offer attractive salaries and a variety of IT-related positions such as database administrator, technical writer, technical support specialist, systems analyst, digital media specialist, web developers/designers, programmers and network administrator, etc. Some IT graduates, who otherwise may not find a typical IT-related job, often venture into more lucrative professions (e.g. business, consultant, project management, etc.) rather than teaching.

While it is important to reiterate Thailand's fiscal and educational policies, we must not neglect the foremost foundations of an educational system—*Kru* and *Ajarn*. We realize that Thai educators work long hours, mentor students, and serve as surrogate parents in some aspects. We also know that IT educators must maintain their expertise by reading recent research or educational literature, collaborating with colleagues, and attending professional training and seminars. As Thailand attempts to strategically reposition its educational system to become a regional leader in Southeast Asia; examining the perspectives of those who are current working in the teaching profession will be necessary.

This study will extend our basic understanding of Thai IT educators by exploring the necessary ingredients that cultivate career satisfaction and teaching commitment through the organizational experiences. The goals of this study will be to apply Igbaria & Wormley's model of organizational experiences [13] in Thai university settings and to examine the relationships between organizational experiences (i.e., organizational acceptance, met expectation, training programs, career support, and job discretion), career satisfaction, and teaching commitment. It also examines the impact of organizational experiences on career satisfaction, which leads to teaching commitment.

2 Theoretical Framework

To examine the vital ingredients for career satisfaction and teaching commitment, Igbaria and Wormley's constructs of organizational experiences were used [13]. These organizational experiences— *organizational acceptance, met expectations,*

training programs, career support and job discretion—are the antecedents to career satisfaction. Shown in Figure 1, these five organizational experiences have been shown to influence *career satisfaction*, which affects *teaching commitment*. Previous studies have examined the relationship between organizational experiences and career satisfaction [6,10,11,14,16].

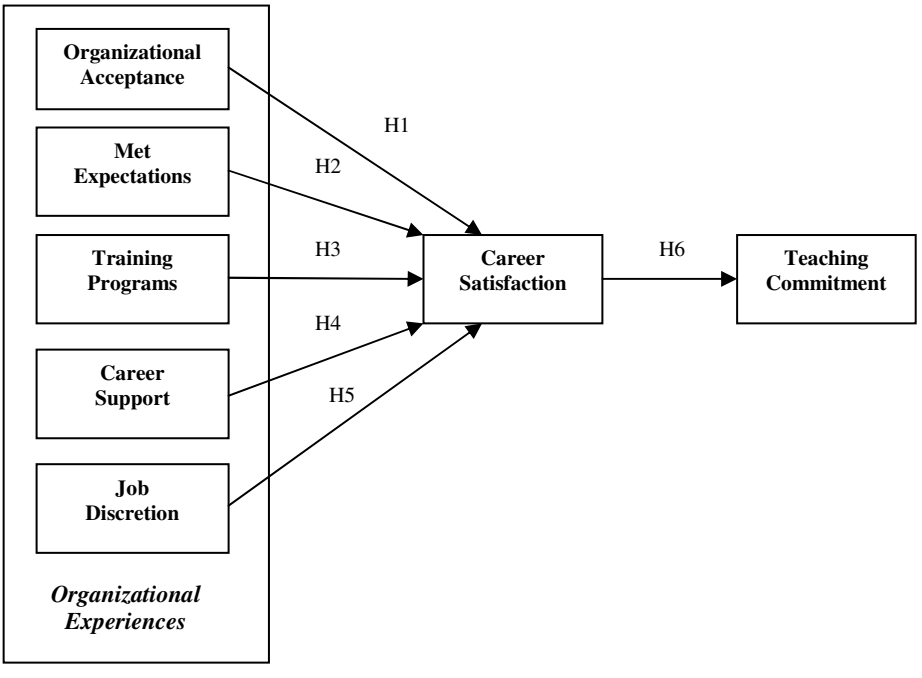


Fig. 1. Proposed Research Model

Prior research finds that organizational acceptance is positively related to career satisfaction through promotion [7]. Reports from Computing Research Association (www.cra.org) put organizational morale/culture as a primary factor in accepting a university's job offer, as new faculty desired schools with greater organizational acceptance.

Hypothesis 1: Organizational acceptance is positively related to career satisfaction.

A longitudinal study shows that met expectations along peer relationships significantly predict organizational commitment, turnover intention, and job satisfaction [15]. Several aspects of newcomer's expectations serve as the antecedents to satisfaction and commitment [21]. Similarly, employees' expectations have been found to impact career outcomes [8].

Hypothesis 2: Met expectations are positively related to career satisfaction.

Having appropriate training programs is one of the key strategies to solve turnover intention among IT professionals [22]. Obviously, active participation in organizational sponsored training will promote job growth and personal achievement. Formal

professional development activities have a positive impact on promotion and organizational commitment [14]. In addition, training programs are positively related to job satisfaction [4].

Hypothesis 3: Training programs are positively related to career satisfaction.

Supervisor support is posited to affect productivity and well-being [20]. A study suggests a possible linkage between career support and career satisfaction when race is taken into account [7]. Employees who receive career-oriented mentoring report higher work commitment and career satisfaction [2].

Hypothesis 4: Career support is positively related to career satisfaction.

The ability to exercise discretion may be positively related career satisfaction. There is a moderate correlation between job autonomy and job satisfaction [3]. Employees who perceive themselves as having less discretion on the job generally have lower performance and career satisfaction [7].

Hypothesis 5: Job discretion is positively related to career satisfaction.

Studies demonstrate the relationships among leadership support, job satisfaction, and organizational commitment [9]. Researchers find a significant correlation between job satisfaction and organizational commitment [5].

Hypothesis 6: Career satisfaction is positively related to teaching commitment.

3 Research Method

3.1 Data Collection

The targeted respondents were Thai postsecondary IT educators (i.e., computer science, information systems, or other IT-disciplines) from vocational schools, colleges

Table 1. Characteristics of the Respondents (N=71)

Characteristics	Percentage
University	
Public	45.7
Private	54.3
Gender	
Male	52.1
Female	47.9
Work Experience (yrs)	
<1	14.1
1-3	23.9
4-6	26.8
7-9	11.3
>9	23.9
Academic Tenure	
Lecturer	91.5
Assistant/Associate	8.5

and universities. A web-based survey was conducted in Thai with English translation for some ambiguous questions. Chairs or heads of twenty-eight IT departments were contacted in order to distribute the survey to their faculty members. A follow-up email was sent to IT department chairs two weeks later stating the goals of this study and benefits that would return to the academic profession and IT educators. Participation in the survey was completely voluntary and confidential. Nearly 560 IT educators received the survey. Seventy-one participants completed and replied to the survey—a response rate of 12.7%. Average age of the respondents was 33.6 years (S.D.=7.3). The table below shows the characteristics of the respondents.

3.2 Measurement Reliability

All of the measurement items were taken from the existing literature [1,7,13,17,18,19] and translated into Thai. With the exception of *training programs (TP)*, the study performed the following activities on all of the constructs: (1) Exploratory Factor Analyses (EFAs) were performed to examine convergent validity of the questionnaire items. The study eliminated any item with significant cross-factor loadings or any item that did not increase the internal consistency of its designated construct. The twelve items from Teaching Commitment (TC) had been converted to Likert-type scales forming a four-item second-order construct based on pay, autonomy, promotion, and relationships. (2) Cronbach's α was calculated for all of the items within the same construct. The final factor analysis with varimax rotation produced a distinctive six-factor solution with a significant factor loading of .65 (p=.05): Organizational Acceptance (α=.893), Met Expectation (α=.565), Career Support (α=.948), Job Discretion (α=.694), Career Satisfaction (α=.908), and Teaching Commitment (α=.830). Appendix A shows the final list of the questionnaire items used to test the hypotheses.

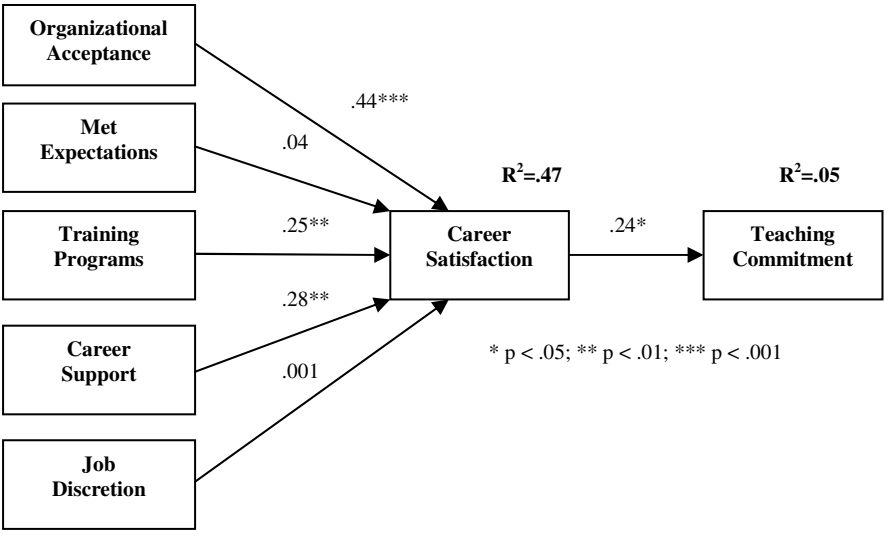


Fig. 2. Results of Linear Regression

3.3 Hypotheses Testing

To test the proposed hypotheses, collinearity diagnostic and linear regression analyses were performed on the research model. The average variance inflation factor (VIF) of the predictor variables was 1.16—a confirmation that collinearity was not a problem when regressing career satisfaction on organizational experiences. Figure 2 illustrates the results of multiple regression analyses along with the standardized coefficients. The result of the first linear regression (F=13.5; p<.001) showed that 47 percent of the variances in career satisfaction were predicted by organizational acceptance (t=4.6; p<.001), training programs (t=2.7; p< .01), and career support (t=2.9; p<.01). Non-significant t-values (p>.05) of met expectation (t=.37; p>.05) and job discretion (t=.01; p>.05) also resulted in the rejection of hypotheses 2 and 5. The result of second linear regression (F=4.37; p<.05) confirmed that career satisfaction significantly predicted teaching commitment (t=2.09; p<.05) with only 5 percent of the shared variance.

4 Discussion and Implications

Organizational acceptance had a positive impact on career satisfaction as hypothesized. This finding is not surprising in a collective society. The department culture and environment will likely influence career satisfaction of IT educators in Thailand. Social bonding activities, especially for part-time lectures, will build rapport among fellow teachers. Peer or senior faculty mentoring continues to be a crucial element in continuous collaboration. Private universities can allow senior lecturers involvement in various decision-making processes.

Met expectations did not lead to career satisfaction as hypothesized. Although it was hypothesized to predict career satisfaction, the original Igbaria and Wormley's model used job performance and advancement prospects as the mediating variables between met expectations and career satisfaction [13]. In this study, the questions of met expectations were related to promotional rate and salary increment. Other sub-components (i.e., current career opportunities, training, and development experiences) of met expectations could alter the result as the majority of our respondents were lecturers with no rank. To increase the reliability of the construct, these sub-components were excluded during the exploratory factor analyses. However, the inflated α of .565 was still below the acceptable value of .60. We rationalize that the elimination of these sub-components and the absence of Igbaria & Wormley's intermediate constructs [13] may actually hinder the predictability of our proposed model.

On the other hand, competitive salaries can reduce turnover intention, but may not be a necessary ingredient for career satisfaction among Thai IT educators, especially when we consider other intangible incentives such as increased social status and respect. With implicit social and moral responsibilities, Thai society often demands greater respect for teachers and other individuals (e.g., parents, teachers, medical doctors, monks, etc.) In a collective and high *power distance* society, a disrespectful gesture toward these individuals is interpreted as an act of disenfranchising oneself from one's own family and society.

Our study found training programs positively related to career satisfaction. Inadequate professional development activities can limit advancement prospects of IT professionals [14]. An opportunity to participate in formal and informal training activities helps promote personal growth and is associated with career success [19]. This opportunity can support full-time IT lecturers towards a terminal degree.

Career support was shown to be positively related to career satisfaction. The finding is consistent with the previous research [7]. Supervising support is a motivating factor in a high *uncertainty avoidance* society. According to a participant's comment, the head of the department or college should support his attempt to acquire additional professional development activities, especially in doing research. Similar to the U.S., promotion through various academic ranks—starting from Assistant Professor—relies on a strong publication record. Unfortunately, many Thai IT educators do not have the time, motivation, or financial incentives to conduct or publish quality research. As a result, only a dozen universities in Thailand are deemed research universities. Several of our participants mentioned opportunities to further their doctoral study. Advanced education, located time, tangible incentives, and personal motivation and commitment can cultivate quality research.

Job discretion was positively related to career satisfaction when job performance served an intermediate variable [7]. However, our study found no relationship between job discretion and career satisfaction. Job discretion—used here as a two-item construct—had an unreliable, inflated α value. To what extent did our respondents exercise job discretion was questionable since organizational decision-making would not be considered as a key factor. The majority of our respondents were lecturers, not academic administrators.

Career satisfaction was positively related to teaching commitment as proposed. Nevertheless, only 5 percent of the variance of teaching commitment was predicted by career satisfaction. In other words, career satisfaction was not the main determinant of teaching commitments among Thai IT lecturers. Since the study measured teaching commitment via salary, job freedom, promotion, and friendliness of coworkers, the instrument might not be applicable to Thai society. As mentioned earlier, Thai IT educators are generally committed to the teaching profession due to various intrinsic motivations. For example, there is a belief that *Kru* or *Ajran* have a self-sacrificing obligation to pass on the knowledge and wisdom to students, whose future prosperity will benefit their country. It is probably this belief, and other existing virtues that influence teaching commitment and help sustain the academic profession in Thailand.

5 Limitations and Future Research

The questionnaires were conducted in Thai with some English translation. The direct translation of English to Thai could be confusing in some instances. Given the small sample size (N=71), the majority of respondents were lecturers. Future research would benefit from the sampling of a more representative sample of Thai IT educators. Since the study did not investigate job performance and advancement prospects [13], which could be more relevant to autonomous state universities in the near future, researchers should consider these variables along with other characteristics such as

professor rank, type of university/college (public versus private), demographic and region setting.

We also gain additional insight from this study. Future research related to organizational experiences, career satisfaction and teaching commitment among Thai educators needs to be reexamined from a socio-cultural perspective. A generic measure of career satisfaction among IT professionals may not applicable to IT educators. In addition, continued exploration and investigation of organizational experiences of career satisfaction of IT educators may be fruitful in longitudinal research—the assessment of seven variables could be measured at different periods of one's career. This avenue of inquiry might serve to identify how realistic IT educators' perceptions of career satisfaction are in the beginning of their careers and how their opinions change over time.

6 Conclusion

This work makes several contributions to the MIS Human Resource Management literature. It is the first to examine career satisfaction and teaching commitment of Thailand's information technology postsecondary instructors. Our findings suggest that Thai IT educators are motivated by many of the same factors as their American counterparts. However, there are important differences. As Thailand is a *collectiv*e and high *uncertainty avoidance* society, it is not surprising that *job discretion* was not significantly related to *career satisfaction*. Our findings also suggest that *organizational acceptance, training programs,* and *career support* are each positively related to *career satisfaction* for Thai IT educators. As such, this work provides important guidance for those tasked with recruiting and retaining Thai IT educators. Our research suggests that much of the advice given by Igbaria and Wormley [13] to supervisors of MIS employees is relevant to supervisors of Thai IT educators. For example, given the importance of *organizational acceptance* by our survey participants, efforts aimed at facilitating the acceptance of IT educators by others in the university may improve recruiting and retention. Similarly, our work suggests that training programs and career support programs (e.g., career mentoring or degree completion initiatives) may improve career satisfaction and ultimately retention.

References

1. Alutto, J.A., Hrebiniak, L.G., Alonso, R.C.: On Operationalizing the Concept of Commitment. Social Forces 51, 448–454 (1973)
2. Aryee, S., Chay, Y.W.: An Examination of the Impact of Career-Oriented Mentoring on Work Commitment Attitudes and Career Satisfaction Among Professional and Managerial Employees. British Journal of Management 5, 241–249 (2005)
3. Blegen, M.A.: Nurses' Job Satisfaction: A Meta-analysis of Related Variables. Nursing Research 42, 36–41 (1992)
4. Chiang, C., Back, K., Canter, D.: The Impact of Employee Training on Job Satisfaction and Intention to Stay in the Hotel Industry. Journal of Human Resources in Hospitality & Tourism 4, 99–118 (2005)

5. Chiang, T., Wang, J.: Workplace Learning, Job Satisfaction, and Organizational Commitment in Small to Midsize Companies in Taiwan. In: Academy of Human Resource Development International Research Conference in the Americas, Panama City, Florida (2008)
6. Egan, T.M., Yang, B.B., Kenneth, R.: The Effects of Organizational Learning Culture and Job Satisfaction on Motivation to Transfer Learning and Turnover Intention. Human Resource Development Quarterly 15, 279–301 (2004)
7. Greenhaus, J.H., Parasuraman, S., Wormly, W.M.: Race Effects on Organizational Experiences, Job Performance Evaluation, and Career Outcomes. Academy of Management Journal 33, 64–96 (1990)
8. Greenhaus, J.H., Seidel, C., Marinis, M.: The Impact of Expectations and Values on Job Attitudes. Organizational Behavior and Human Performance 31, 394–417 (1983)
9. Hulpia, H., Devos, G., Rosseel, Y.: The Relationship between the Perception of Distributed Leadership in Secondary Schools and Teachers' and Teacher Leaders' Job Satisfaction and Organizational Commitment 20, 291–317 (2009)
10. Igbaria, M., Greenhaus, J.H.: Determinants of MIS Employee' Turnover Intentions: A Structural Equation Model. Communications of the ACM 35, 34–49 (1992)
11. Igbaria, M., Guimaraes, T.: Antecedents and Consequences of Job Satisfaction among Information Center Employees. Journal of Management Information Systems 9, 145–174 (1993)
12. Igbaria, M., Guimaraes, T.: Exploring Differences in Employee Turnover Intentions and Its Determinants among Telecommuters and Non-Telecommuters. Journal of Management Information Systems 16, 147–164 (1999)
13. Igbaria, M., Wormley, W.M.: Organizational Experiences and Career Success of MIS Professionals and Managers: An Examination of Race Differences. MIS Quarterly 16, 507–529 (1992)
14. Mahatanankoon, P.: The Effects of Post-Educational Professional Development Activities on Promotion and Career Satisfaction of IT Professionals. In: Proceedings of the 2007 ACM SIGMIS-CPR Conference on Computer Personnel Research, pp. 9–14. ACM, New York (2007)
15. Major, D.A., Kozlowski, W.J., Chao, G.T.: A Longitudinal Investigation of Newcomer Expectations, Early Socialization Outcomes, and the Moderating Effects of Role Development Factors. Journal of Applied Psychology 80, 418–431 (1995)
16. Nelson, R.R., Todd, P.A.: Strategies for Managing IS/IT Personnel. IGI Publishing, Hershey (2004)
17. Nixon, R.: Black Mangers in Corporate America: Alienation or Integration. National Urban League, Washington (1985)
18. Nixon, R.: Perceptions of Job Power among Black Managers in Corporate America. National Urban League, Washington (1985)
19. Schambach, T., Blanton, J.E.: The Professional Development Challenge for IT Professionals. Communications of the ACM 45, 83–87 (2002)
20. Turner, J.A., Baroudi, J.J.: The Management of Information Systems Occupations: A Research Agenda. ACM SIGCPR Computer Personnel 10, 2–11 (1986)
21. Wanous, J.P.: Organizational Entry: Recruitment, Selection, and Socialization of Netcomers. Addison-Wesley, Reading (1980)
22. Wingreen, S.C., Blanton, J.E., Kittner, M.L.: The Relationship between IT Professionals' Individual Factors, Training Climate Fit, and Turnover Intentions. In: Proceedings of the 2002 ACM SIGCPR Conference on Computer Personnel Research, pp. 163–166. ACM, New York (2002)

Appendix A: Final Questionnaire Items Used

Organizational Acceptance
1. There is a feeling of comraderie between me and my work associates
2. I like the people with whom I work
3. I am really a part of my work group

Training Programs
1. Attending professional conferences
2. Attending professional/technical seminar
3. College IT courses (non-degree seeking)
4. In-house training supplied by employer
5. Teaching courses (or preparing material)
6. Discussing new technologies with peers
7. Reading professional newsletters/magazines
8. Reading professional journals/books
9. Using e-mail to consult with networks of peers

Job Discretion
1. I have very little responsibility on my job
2. I have considerable decision making power on my job

Teaching Commitment
1. No increase in pay
2. Slight increase in pay
3. Large increase in pay
4. No increase in job freedom
5. Slight increase in job freedom
6. Large increase in job freedom
7. No increase in status
8. Slight increase in status
9. Large increase in status
10. No increase in friendliness of co-workers
11. Slight increase in friendliness of co-workers
12. Large increase in friendliness of co-workers

Met Expectations
1. My rate of promotion has been…
2. My salary increases over the years have generally been…

Career Support
1. My supervisor takes the time to learn about my career goals and aspirations
2. My supervisor cares about whether or not I achieve my career goals
3. My supervisor keeps me informed about different career opportunities for me in the organization
4. My supervisor makes sure I get the credit when I accomplish something substantial on the job
5. My supervisor gives me helpful feedback about my performance.
6. My supervisor gives me helpful advice about improving my performance when I need it
7. My supervisor supports my attempts to acquire additional training or education to further my career
8. My supervisor provides assignments that give me the opportunity to develop and strengthen new skills
9. My supervisor assigns me special projects that increase my visibility in the organization

Career Satisfaction
1. I am satisfied with the success I have achieved in my career
2. I am satisfied with the progress I have made toward meeting my overall career goals
3. I am satisfied with the progress I have made toward meeting my goals for income
4. I am satisfied with the progress I have made toward meeting my goals for advancement
5. I am satisfied with the progress I have made toward meeting my goals for the development of new skills

Knowledge-Centric Management of Business Rules in a Pharmacy

Juha Puustjärvi[1] and Leena Puustjärvi[2]

[1] Helsinki University of Technology, Box 9210, 02015 TKK, Finland
juha.puustjarvi@tkk.fi
[2] The Pharmacy of Kaivopuisto, Neitsytpolku 10, Helsinki 00140, Finland
leena.puustjarvi@kolumbus.fi

Abstract. A business rule defines or constraints some aspect of the business. In healthcare sector many of the business rules are dictated by law or medical regulations, which are constantly changing. This is a challenge for the healthcare organizations. Although there is available several commercial business rule management systems the problem from pharmacies point of view is that these systems are overly geared towards the automation and manipulation of business rules, while the main need in pharmacies lies in easy retrieving of business rules within daily routines. Another problem is that business rule management systems are isolated in the sense that they have their own data stores that cannot be accessed by other information systems used in pharmacies. As a result, a pharmacist is burdened by accessing many systems inside a user task. In order to avoid this problem we have modeled business rules as well as their relationships to other relevant information by OWL (Web Ontology Language) such that the ontology is shared among the pharmacy's applications. In this way we can avoid the problems of isolated applications and replicated data. The ontology also encourages pharmacies business agility, i.e., the ability to react more rapidly to the changes required by the new business rules. The deployment of the ontology requires that stored business rules are annotated by appropriate metadata descriptions, which are presented by RDF/XML serialization format. However, neither the designer nor the pharmacists are burdened by RDF/XML format as there are sophisticated graphical editors that can be used.

Keywords: Ontologies, Business rules, Knowledge management, Business agility, Knowledge centric organization.

1 Introduction

The function of pharmacies is to ensure the safe and effective use of pharmaceutical drugs. As a result, the scope of pharmacy practice is wide: it includes traditional roles such as compounding and dispensing medications, reviewing medications for safety and efficacy, and providing drug information [1]. Further pharmacists are expected to be the experts on drug therapy and primary health professionals, who optimize medication use to provide patients with positive health outcomes.

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 116–125, 2010.

Ensuring to maintaining pharmacists' expertise requires large efforts as healthcare is a field where the fast development of drug treatment and the introduction of new drugs require specialized skills and knowledge that need to be renewed frequently [2, 3]. As a result, also the amount of new information concerning new medication increases rapidly.

Pharmacies receive medicinal information from a variety of sources [4], e.g., from medical authority, medicinal wholesalers, and pharmaceutical companies. These information entities arrive in variety formats, e.g., by paper mail, e-mail, and fax. Also the nature of the information entities may vary, e.g., an information entity may be a learning object, a regulation, a guide or a bulletin [5]. Further, some of the information gives rise for a new business rule or changing prevailing rules.

Business rules exist in pharmacies - as well as in any organization - although they are not necessary written down, talked about or even part of the organization's consciousness. Business rules define or constraint some aspect of the business [6]. They are means by which organization's strategy is implemented. They attempt to describe the prevailing practices of an organization, but they are not an attempt to specify how the organization should work [7].

In health care sector business rules are usually gathered only when they are dictated by law, or when they are required for automating organization's processes [8]. However, business rules should be gathered and presented explicitly as they comprise a critical asset of an organization. They can be exploited in many ways including:

- to automate business processes,
- to re-engine business processes,
- to guide new employees,
- to demonstrate the fulfillment of legal obligations, and
- to communicate among organizations principals and third parties.

Business rule management systems are specialized for business rule management, automation and re-engineering [9]. Such systems have their own data stores and user interfaces and they provide means for designing and editing rules, which are then automatically enforced by workflow engines. However, a lack of these systems is that they are isolated in the sense that they do not provide a means for accessing other information entities. In particular, this lack appears when a user accesses business rule and other related information in the same user task.

Storing and organizing business rules in a way that they are easily accessible through their connections to other information entities require the introduction of a shared data store, which captures business rules and other relevant information entities. In this paper we have focused on this issue. Our key idea is to model business rules as well as their relationships to other relevant information by OWL (Web Ontology Language) [10] such that the ontology is shared among the organization's applications. In this way we can avoid the problems of isolated applications and replicated data.

The rest of the paper is organized as follows. First, in Section 2, we give a motivating example which illustrates how business rules and other information entities can be interweaved in a user task. The example suggests that the systems dedicated only for managing business rules are not appropriate in pharmacies. In Section 3, we characterize the business rules discovered from pharmacies. Then, in Section 4, we consider metadata and taxonomies as well as their usability in managing business rules and

other related information entities. In Section 5, we first consider the gains of knowledge centric organization and the role of knowledge base. After that we present the pharmacy ontology, which is stored in the knowledge base. Finally, Section 6 concludes the paper by discussing the advantages and disadvantages of our introduced solutions.

2 Motivating Example

Pharmacies as well as any organization receive digital documents from a variety of external sources. Depending on the nature of the received document and the deployed systems, the document is stored in an appropriate system, which may be for example:

- Content Management system,
- Learning Content Management System,
- Document Management Systems,
- Database system
- Customer Relationship Management Systems, or
- Business Rule Management System.

Traditionally such systems have their own isolated data stores and user interfaces. As a result, if the data needed in a user task is not included in one system, then the user has to log and access data from various systems. Further if the systems do not support SSO (Single Sign-On) this may require many user activities. By SSO we refer to the technical mechanism that allows the user to only authenticate once to her client, so that she does not have to memorize several usernames and passwords when accessing more than one system.

To illustrate the use of two systems in a user task in a pharmacy let us assume that a pharmacist is dispensing a medicinal product, say Diovan, for a customer. From customers social insurance card the pharmacist recognizes that the patient is a veteran. The pharmacist remembers that the discount of ten per cent is granted for veterans of some drugs, but she does not remember whether it concerns Diovan. By accessing the business rule management system by the keywords “discounts” and “veterans” the system return the statement “discount of 10 per cent of the prescription based partially repayable drugs is granted for veterans”. Next the pharmacist opens and finds from the medicinal system that Diovan is not a partially repayable drug. Then she retrieves from the pricing system the medicinal products that are substitutable with Diovan and are partially repayable. Finally, by the permission of the client the pharmacist change Diovan to a substitutable medicinal product and grants the discount of 15 per cent for the customer.

Here the reason for the complexity of the pharmacist user task is that the data needed by a pharmacist is stored in separate systems, and their data cannot be accessed by other systems. In principle, this problem can be avoided by integrating the data stores into a knowledge base and by sharing the integrated data store among the systems. However, an interesting question arising now is how business rules and other relevant informal material should be presented, organized and retrieved. The representation and organization of business rules should provide the pharmacist with an easy access to the business rules and other relevant information.

3 Business Rules

Business rules are seldom stated explicitly and centrally in a pharmacy. Rather they are disseminated throughout the organization: they can be stated in health government regulations received from authorities or in business contracts where they are represented by terms and conditions of the contract. Business rules can also exist in the form of common knowledge among the employees.

Business rules can be expressed in a natural language, in a graphical way or in a formal language [6]. Which expression formats are appropriate depends on the nature of the rule and the way the rule is used, e.g., whether the enforcement of the rule is automated by a database system, workflow engine or business rule management system. Business rule management system is a software system that helps managing business rules in a central rule repository, and so they provide additional functionality to maintain the rules [9]. They also execute rules in a dedicated rule engine. In particular they externalize the rules and so business rules can be easily modified.

The gain of using a natural language is that employees can easily express and understand the rules. Hence, business rules should be expressed also in a natural language even though they were automated.

The gain of using graphical notations is that the long and difficult textual descriptions can be avoided. For example, a business rule, which determines a refund of specific drug for specific customer group requires long and difficult textual descriptions.

The gain of formal rules is that semantic reasoners [11] can be used in discovering the contradicting rules as well as in deciding the validity of the business rules. However, exploiting semantic rules requires the introduction of a business rule management system that supports reasoning. The introduction of a rule system in small organizations like pharmacies is not always appropriate as they require rather high investments on software and training.

Business rules are commonly classified into four categories: business terms, facts, constraints and derivations [6]. This classification also suits well for the rules discovered from pharmacies.

- *Business term* is a word or phrase that has a specific meaning for a business in some designated context. They are usually presented in a glossary or as the entities in a conceptual model. For example, *drug, medicinal product*, *dose* and *dosage* represent business terms in a pharmacy.
- *Facts* relate business terms to each other. For example, statement "*each medicinal product includes one or more drugs*" is a fact. Facts can be presented by a natural language or by conceptual models.
- *Constraints* are used to determine whether the results of a transaction or operation are correct. For example a constraint in a pharmacy may state that "*When the amount of the medicinal product Diovan is less than 8, then an order 20 units is shipped to the wholesaler*". This kind of rules can be automated business rule management system or by inventory systems by presenting them as ECA (Event Condition Action) rules [12].

- *Derivation rules* transform received information into returned values. For example, in pharmacy the sickness number in a social insurance card determines the discount to be granted for the customer, e.g., the number corresponding veterans gives rise for ten percent discount.

After the business rules gathered they should be stored and linked to the right activities. Those business rules which enforcement is automated are implemented in program code or by business rule management system. A drawback of program code based implementation is that the modifications as well as the implementation is time consuming and cannot be easily done as the rules are hardcoded in applications [6]. These problems can be avoided by using business rule management systems.

4 Attaching Metadata to Business Rules

4.1 Business Rule Metadata

Metadata is data about data [13]. It is intended to facilitate the discovery of electronic resources from the Web. Metadata describes certain important characteristics of its target, e.g., a document representing a business rule.

We make the distinction between syntactical and semantic metadata. *Syntactical metadata* describes the structural characteristics of its target, such as the format, language, date, creator, and the author of the document. Dublin Core [14] is a widely used metadata standard that represents syntactical metadata. Although syntactical metadata standards are useful in managing business rules and other information entities they do not enable content based retrieval and so semantic description are also needed.

Semantic metadata describes the semantic content of the target [15]. For example, the domain specific keywords, attached to documents represent semantic metadata. For example, *painkiller* could represent a semantic metadata of a *drug*. Further, in order to standardize the used semantic metadata items certain domain specific taxonomies and ontologies are needed.

We next consider appropriate taxonomies and ontologies for standardizing business rule metadata items discovered from pharmacies.

4.2 Business Rule Taxonomy

Taxonomy is a way to classify or categorize a set of things into a hierarchy [16]. The logic behind taxonomy is that when one goes up the taxonomy toward the root, the information entities become more general, and respectively when one goes down towards the leaves the information entities become more specialized. We can also state this in a more formal way: depending on the direction of the link each link between a parent and a child node represents a subclassification relation or superclassification relation. For example, *painkiller* is a subclassification of *drug*, and *drug* is a super-classification of *painkiller*.

Applying taxonomies in retrieving business rules requires that each business rule is augmented by a metadata items that are taken from one or more taxonomies. A pharmacist can then query business rules by Boolean expressions comprising of operands

and operations. The operands are the used keywords (nodes in the taxonomy tree) and the operands are typically “and”, “or”, and “not”.

The idea behind our developed taxonomies is to connect daily routines into appropriate documents such that in processing a user task, which may require data retrieval, the used keywords can be taken from one taxonomy. As a result we have developed taxonomies such as *customer taxonomy*, *medicinal product taxonomy*, *pricing taxonomy*, *procurement taxonomy*, *inventory taxonomy* and *medicinal information taxonomy*. To illustrate this, a subset of the customer taxonomy is presented in Figure 1.

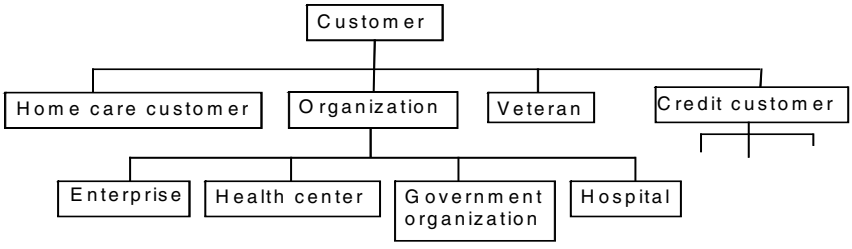


Fig. 1. Customer taxonomy

The Boolean model is intuitive and clear. However there is wide variety of queries that cannot be expressed by keyword such as the motivating example given in Section 2, where the pharmacist retrieved the medicinal products that are substitutable with medicinal product Diovan.

In order to be able to execute such queries there must be a conceptual schema or ontology that models the relationship of the concepts (terms). Such a feature is included in our designed pharmacy ontology. In addition as we will illustrate the ontology captures the used taxonomies and business rules’ metadata, and so it also provide a means for supporting metadata (keyword) based searching of business rules.

5 Knowledge-Centric Organization and Pharmacy Ontology

5.1 Knowledge-Centric Organization

Knowledge management concerns with acquiring, accessing and maintaining knowledge within an organization [16]. The key idea in knowledge centric organizations is to revolve all applications around the shared ontology [17]. In our case, as illustrated in Figure 2, it means the integration of the data repositories of the pharmacy’s systems such as business rule management system, content management system and learning object management system. To the integrated ontology we refer by the term *pharmacy ontology*. So the pharmacy’s systems can seamlessly interoperate through accessing the shared pharmacy ontology.

A subset of the pharmacy ontology is graphically presented in Figure 3. In the figure ellipses represent classes and boxes represent properties.

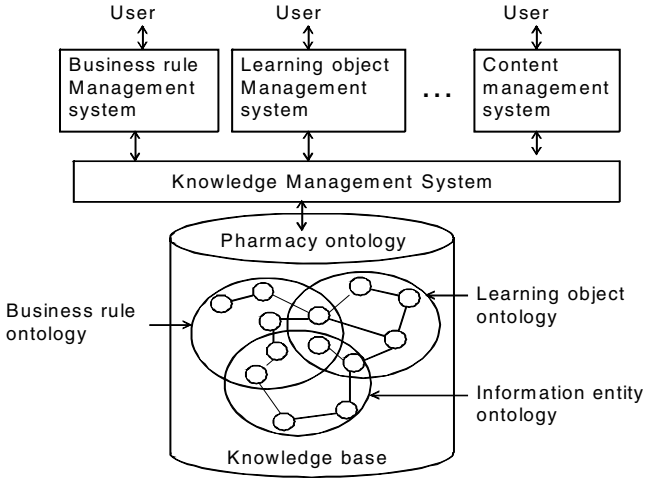


Fig. 2. The knowledge centric architecture of pharmacy's systems

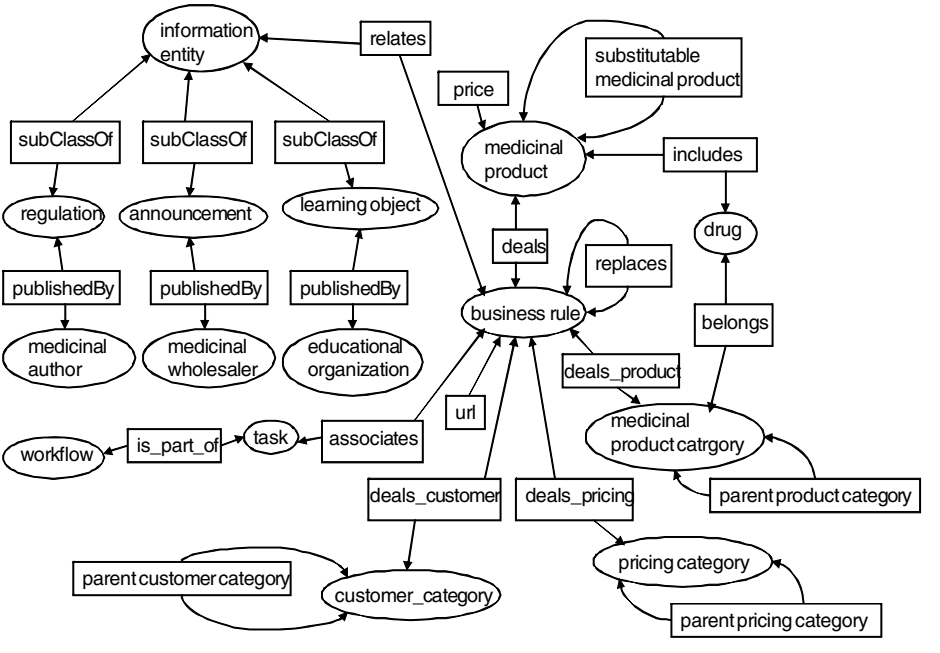


Fig. 3. A subset of the pharmacy ontology

Among other things the pharmacy ontology models business rules as well as their relationships to other relevant concepts such as information entities, which may be regulations, announcements or learning objects. The ontology also models medicinal product as well as the substitution of medicinal products by other (cheaper) medicinal products.

The pharmacy ontology enables queries like the one in the motivating example given in Section 2, where the pharmacist requested the medicinal products that are substitutable with the medicinal product Diovan. In addition, the ontology models the associations of business rules to the tasks of the day-to-day work. Hence, the business rules that are specific to certain tasks can be easily retrieved. The pharmacy ontology also encourages pharmacy's business agility, i.e., pharmacy's ability to react more rapidly to the changes required by the new business rules.

As illustrated in the graphical ontology a *url* (Universal Resource Locator) is an attribute of the class *business rule*. This means, that the explicit business rules are not stored in the ontology but rather the location where they are actually stored. This simplifies the management the rules as they may be located for example in the document archives of health care authorities.

The pharmacy ontology also includes the taxonomies *Customer category*, *Pricing category* and *Medicinal product category* as well as the relationships of the business rules into these categories. Hence, the ontology captures the metadata items attached to business rules as well as the taxonomies from which the keywords are taken. Note that the taxonomies are modeled by specifying the parent of each node in the taxonomy. As the root has no parent, it references to the node itself.

5.2 Representing the Pharmacy Ontology in OWL

By an ontology language it is possible to write explicit, formal conceptualizations of domains. OWL (Web Ontology Language) [18] is an ontology language, which has well defined syntax, a formal semantics and efficient reasoning support.

The instances of the OWL ontologis are presented by RDF [19], which itself is a data model. Its modeling primitive is an object-attribute-value triple, which is called a statement. A description may contain one or more statements about an object. RDF-statements are usually coded by XML, i.e., they are presented in XML-serialization format [20].

OWL-ontologies are also represented by RDF i.e., they are comprised of RDF-statements. Thus we can present queries on the pharmacy ontology by query languages developed for RDF, e.g., by SPARQL [21]. It is standardized by the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium, and is considered a component of the semantic web.

6 Conclusions

The fast development of drug treatment entails a huge amount of educational and informal learning material that has to be disseminated for the pharmacies and further to their pharmacists. Part of the incoming information gives rise for new business rules or for changing the prevailing ones. As many business rules in pharmacies are dictated by law or medical regulations pharmacies should be able to react rapidly to the changing business rules. This in turn requires that pharmacists should have easy access to business rules in their daily routines.

On the other hand, business rules represent just one facet of the information that pharmacists have to access. Other actively used information includes guides, learning

objects and regulations. From daily work-pattern's point of view all the relevant data inside a user task should be accessible though one system. Thereby business rule management systems, which only provide means for automating and manipulating business rules are not appropriate for pharmacies. Instead integrated systems that would provide a way for accessing all relevant information would be of high importance.

We have introduced the idea of knowledge centric organization for managing business rules and other relevant information in a pharmacy. Its corner stone is the pharmacy ontology, which models business rules and other relevant information. The pharmacy ontology also encourages pharmacies business agility, i.e., pharmacies ability to react more rapidly to the changes required by the new business rules.

Storing business rules to the ontology requires that they are annotated according to the ontology. This is an extra effort required by the introduction of the pharmacy ontology. However, though the insertions are stored in RDF/XML serialization format neither the designer nor the user are burdened by RDF/XML format as there are sophisticated graphical editors that can be used in annotating the rules.

Although our approach is geared towards the information retrieval rather than automation of business rules our approach also contributes to the automation of business rules as they are organized and classified in a way that designer do not have discover business rules before their automation.

References

1. Batenburg, R., Van den Broek, E.: Pharmacy information systems: the experience and user satisfaction within a chain of Dutch pharmacies. International Journal of Electronic Healthcare 4(2), 119–131 (2008)
2. Puustjärvi, J., Puustjärvi, L.: Automating the Dissemination of Information Entities to Healthcare Professionals. In: Papasratorn, B., et al. (eds.) IAIT 2009. CCIS, vol. 55, pp. 123–132. Springer, Heidelberg (2009)
3. Puustjärvi, J., Puustjärvi, L.: Integrating Medicinal Learning Objects with Daily Duties. In: Mohammed, S. (ed.) Ubiquitous Health and Medical Informatics, IGI-Global (2010)
4. Khoumbati, K., Shah, S., Dwivedi, Y.K., Shah, M.H.: Evaluation of investment for enterprise application integration technology in healthcare organisations: a cost-benefit approach. International Journal of Electronic Healthcare 3(4), 453–467 (2007)
5. Puustjärvi, J., Puustjärvi, L.: Managing personalized and adapted medical learning objects. In: The Proc. of the 7th IEEE International Conference on Advanced Learning Technologies, ICALT (2007)
6. The Business Rules Group: Defining Business Rules – What are they really? Report, The Business Rules Group (2000)
7. Ross, G.R.: Principles of the Business Rule Approach. Addison-Wesley Publishing Company, Reading (2003)
8. Huseman, S., Schäfer, M.: Building flexible eHealth processes using business rules. In: Proc. of the 1st European Conference on eHealth (ECEH 2006), pp. 131–142 (2006)
9. Chappel, O.: Keeping Business Rules Separate from their Enforcement. Business Rules Journal 6(7) (2005)
10. OWL – WEB OntologyLanguage, `http://www.w3.org/TR/owl-features/`
11. Antoniou, G., Harmelen, F.: A semantic web primer. The MIT Press, Cambridge (2004)

12. Ullman, J., Widom, J.: Principles of Database Systems. Prentice Hall, Englewood Cliffs (1998)
13. Puustjärvi, J.: The role of metadata in e-learning systems. In: Ma, Z. (ed.) Web Based Intelligent e-Learning Systems: Technologies and Applications. Idea Group Inc., USA (2005)
14. Dublin Core, http://www.dublin.core.org (last accessed 2010-3-30)
15. Puustjärvi, J.: Syntax and Semantics of Learning Object Metadata. In: Harman, K., Koohang, A. (eds.) Learning Objects: Standards, Metadata, Repositories, and LCMS, Informing Science Press (2007)
16. Daconta, M., Obrst, L., Smith, K.: The semantic web.:A Guide to the Future of XML, Web Services, and Knowledge Management. John Wiley & Sons, Chichester (2003)
17. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. In: Padua workshop on Formal Ontology (March 1993)
18. Davies, J., Fensel, D., Harmelen, F.: Towards the semantic web: ontology driven knowledge management. John Wiley & Sons, West Sussex (2002)
19. RDF – Resource Description Language, http://www.w3.org/RDF/ (last accessed 2010-3-30)
20. Harold, E., Scott Means, W.: XML in a Nutshell. O'Reilly & Associates, Sebastopol (2002)
21. SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/ (last accessed 2010-3-30)

Head Pose Estimation on Eyeglasses Using Line Detection and Classification Approach

Pisal Setthawong and Vajirasak Vannija

School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
51500701@st.sit.kmutt.ac.th, vajirasak@sit.kmutt.ac.th

Abstract. This paper proposes a unique approach for head pose estimation of subjects with eyeglasses by using a combination of line detection and classification approaches. Head pose estimation is considered as an important non-verbal form of communication and could also be used in the area of Human-Computer Interface. A major improvement of the proposed approach is that it allows estimation of head poses at a high yaw/pitch angle when compared with existing geometric approaches, does not require expensive data preparation and training, and is generally fast when compared with other approaches.

Keywords: Head Pose Estimation, Human Computer Interface, Image Processing, Eyeglasses, Geometric Model.

1 Introduction

Head pose estimation is a common human task that aims to finding the orientation and pose of the human head. The head pose is an important part of the non-verbal communication method that allows people to figure who the intention of the communication is focused upon. In addition to being an important part of non-verbal communication, in the area of Human-Computer Interaction, head pose data could be used as an alternative or potential input device used to perform tasks on the computer.

Though the process of figuring head pose is considered easy for human thinking, the process of building computer vision systems to find the head pose estimation from images and videos is considered to be a difficult process due to a myriad of challenges. Issues such as lighting, camera distortion, and accessories (e.g. eyeglasses) contribute to the difficulty of creating head pose estimation systems when computing on pixel based information. Over the years, a myriad of head pose estimation systems have been proposed which attempts to solve the head pose estimation system by using different approaches with their own set of assumptions which provides at the end, different sets of advantages and disadvantages.

In this paper, a head pose estimation system is proposed in which assumes that the subject is wearing eyeglasses. The proposed system uses a novel geometric approach that utilizes unique features of eyeglasses to estimate the head pose of the human subject. This approach is beneficial compared to existing methods because it can detect head poses at higher yaw and pitch angles than other geometric approaches, does not require expensive data preparation, extensive training, and is generally computationally fast when compared with similar approaches.

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 126–136, 2010.

2 Existing Head Pose Estimation Methods

There are numerous head pose estimation systems [15] that have been proposed over the years. Some of the early approach is to use templates to approach the problem. The appearance template methods [17] is one approach that utilizes image-based comparison to attempt to match between the subject pose with an existing set of patterns with its existing pose labels. This approach suffers from efficiency issues due to large corpus requirement, and the flawed assumption that different image space can be compared to similarity in poses. A more advanced template based approach is the detector array method [9] that expands the frontal face detection by training multiple face detectors of discrete poses. This approach is beneficial to the appearance template methods as it does not require a separate head detection and localization step and can ignore some appearance variation that does not correspond to pose changes. The disadvantage of detector array method approaches is that it is expensive to train and prepare many detectors for each discrete pose, and classification issues will arise when the number of detectors is increased.

Another approach is to use nonlinear regression methods [14] that attempts to find non-linear functional mapping from sets of image poses. This approach requires cropped faces that are labeled for training which may suffers from a certain degree of error due to shift, scale, and distortion from the face data that is prepared.

Another alternative approach is to lower the number of dimension spaces in the data. Approaches such as manifold embedding methods [20] assumes there are fewer dimensions in which pose can cause variation. The low-dimension data could then be used for head pose estimation with regression or template matching. The challenge of this method is that it is difficult to separate the head pose while ignoring variations caused by other factors, for example the pose and identity data may be modeled together with this approach.

Approaches that utilize facial features and structures are also popular. One approach in this classification is the flexible model [2][3][25] that attempts to fit a non-rigid model to conform to the facial structure of the subject image. Another approach in this classification is the geometric method [8][16][24] that uses the head shape and configuration of local features to estimate the head pose of the subject. The geometric approach is generally fast and requires only a few features to be detected such as the nose, mouse, and ear to estimate the pose. The drawback of this system comes with the difficult of detecting the features of the face with a high degree of accuracy. Typical feature approaches will not work properly if the subject is facing at a high yaw or pitch angle, or/and has accessories such as glasses that block important features.

One other major approach in head pose estimation is the usage of metrics over a range of frames such as the tracking [26] approach that uses temporal values and movement to estimate the head pose. This approach requires an initial known head pose when the system is initialized or when the subject tracking is lost.

3 Eyeglass Patterns

This section explores typical eyeglass components and their related patterns that could be of useful when used in head pose estimation. In typical geometric approaches, the

usages of facial features are used, whereas accessories such as eyeglasses are considered as artifacts that would cause the system to miss the detection of key features. Based on the assumption that the subject is wearing eyeglasses, we propose a new approach that could be used to estimate the face pose and solve some of the outstanding issues such as detection of poses at higher yaw/pitch pose where facial features used by existing geometric approaches would be obscured.

Typical modern eyeglasses consist of a few parts. The lenses are housed inside a frame. The frame itself would consist of multiple parts such as the nose bridge/pad and legs.

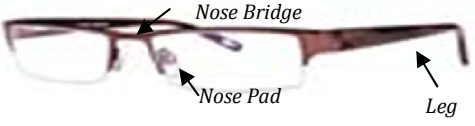


Fig. 1. Eyeglass Parts

One of the important characteristics of eyeglasses on human subject is that when it is worn, the legs of the eyeglasses generally run in a parallel direction to the facing direction of the subject. Based on that pattern, the legs is a candidate for being a feature that could be used to estimate the approximate facing direction and pose of the subject based on the characteristics of the leg contour.

There are many approaches that could be used to detect the leg position of eyeglasses on subject. One approach would be to trace the contour [11] of the eyeglass and figure the leg position from the contour of the eyeglass. As the goal of the system does not require the accurate contour of the eyeglass, but the approximate shape of the leg, it is possible to use simpler approaches to detect the leg of the glasses. The legs of the glasses are typically straight in nature. Due to the nature of legs which are straight, using a line matching algorithm could potentially be a good fit if used to determine if the leg region of the glasses exist [7]. Based on that, it is possible to attempt to detect the legs by trying to detect line-like patterns with the general edge map of the subject.

To generate an edge map, an edge detection algorithm is chosen. As the color of the eyeglasses and skin colors are usually distinctly different, edge detection algorithms will usually generate edge between the regions of the skin and eyeglasses at higher threshold values whereas other natural facial feature features contribute less as they are generally less dominant when a higher threshold value. This behavior is shown in Fig. 2. This assumption makes it possible to quickly prune out edges that are likely not to belong to the contour of the glasses.

Once the edge map has been built, it is possible to attempt to detect lines in the generated edge map to deduce where the eyeglass legs are. Line detection algorithm usually does not detect many lines in the human face as there are few facial features that are linear which follows the initial assumption. However in frontal poses with slight variation, it would be difficult to detect the eyeglass legs as it would be slightly perpendicular to the camera, making the lines formed by the legs to be less dominant. However at higher facing angles, the legs are more dominant and long making it possible for the line detector to easily detect the leg position and orientation.

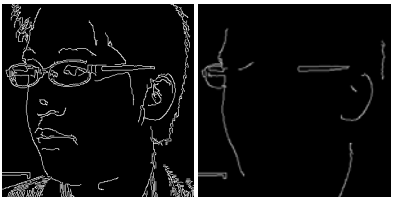


Fig. 2. Increasing threshold of Canny Edge Detector removes much of the facial details from the subject, but retains much of the eyeglass details

4 Proposed System

The proposed system aims to classify images into one of the 9 discrete face pose consisting of Up, UpRight, Right, DownRight, Down, DownLeft, Left, UpLeft, or the Straight pose by using the geometric approach and the newly proposed feature of the eyeglass legs. This system comes with some assumptions. The first assumption is that the subject is wearing eyeglasses. The second assumption is that the proposed approaches allows up to -/+ 90 degrees on the yaw and pitch axis, whereas fixing the roll to simplify the system as the goal is to test the feasibility of using the newly proposed feature in the geometric approach.

4.1 Implemented System Overview

The implemented system was implemented with C/C++ and OpenCV [18] and consists of four major parts and is illustrated in Fig. 3. The first part of the system consists of the Data Preparation components which detects the subject face and prepares the image for later stages of the system. The second component of the system is the Eyeglass Leg Detection component, in which aims to find the eyeglass leg. The third component in the system is the Region Classifier which provides region information for additional processing. The last component of the system is the Head Pose Estimator is used to convert the eyeglass leg data, using additional region information, into one of the discrete pose.

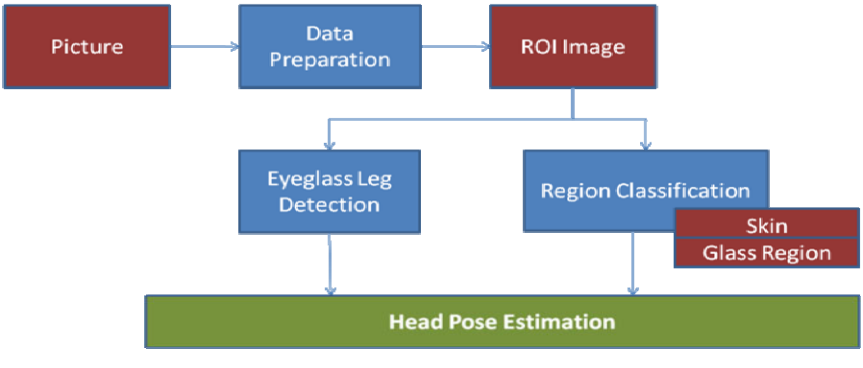


Fig. 3. Overview of Proposed System

4.2 Data Preparation

The Data Preparation stage consists of the Image Acquisition, Face Detection, and Region of Interest (ROI) setup components.

The picture or image is taken from one of the image capture devices through the Image Acquisition process. Once the image is taken, the Face Detection algorithm is used to detect the presence of faces in the picture. Our system uses a face detector based on skin color information blobs [12][22].

Once the face region has been detected, the ROI can be defined for the image. However as parts of the eyeglasses may potentially be outside the detected face area at certain viewing angles, the ROI region is expanded based proportionally on the width and height of the region so that segments of the eyeglasses would not be prematurely be cropped by the ROI region.

4.3 Eyeglass Leg Detection

Once the ROI region has been defined by the system, the next step is to attempt to detect the presence of the eyeglass leg. This process consists of the Edge Detection, Line Detection, Line Pruning, and Line Classification process. The whole process is shown in Fig. 6.

The Edge Detector process is implemented based on the Canny Edge Detector. The threshold of the edge detector used in this system is set to be relatively high. The rationale for high threshold is that the edge detector can prune early many of the edges that are formed by facial features which are weak when compared while retaining stronger edges that are formed by eyeglasses, especially the leg region.

The system initially starts with a threshold of 220 units and attempts to find lines that suit the criteria. However if none is detected, the system then automatically decrements with a fixed step until the threshold reaches 150 units, which is selected based on preliminary experimentation with a sample dataset to be the value where more edges formed are less likely to happen because of the eyeglass legs.

Once the edge information is calculated, the line detector is run on the resulting edge map. The Hough's Line Transform Algorithm [21] was selected to detect candidate lines as the approach can find lines from images and is tolerant of image noise and gaps which makes the algorithm a good fit for this purpose. The parameters to the line detection algorithm are proportionally based upon the parameters of the detected face ROI. Lines detected in this process should propose candidate lines that are potentially the legs of the eyeglasses. Based on experimentation with eyeglass length to the relationship of the size of the ROI image, when considering the pitch/ yaw which is not close to the frontal profile, the eye glass length is between 15% or more of the ROI dimension. Setting the Hough's Line Transform to find lines relative in that range will likely pick up leg candidates that are most probably formed by legs at higher angle and is an effective early pruning process.

The next step is the line pruning process. This process aims to prune candidate lines that are not potentially from the eyeglass legs. Examples include patterns likely formed by background patterns and facial contours which may contribute candidate lines. Region information is used to help in this process and the process is displayed in Fig. 4.

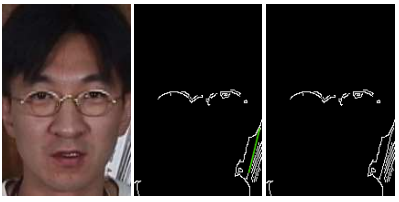


Fig. 4. Example of Line Pruning in action from Original Image, Lines Detected (in Color Before Pruning) Transposed with Edge Map which are caused mostly by background patterns, and Resultant Pruning.

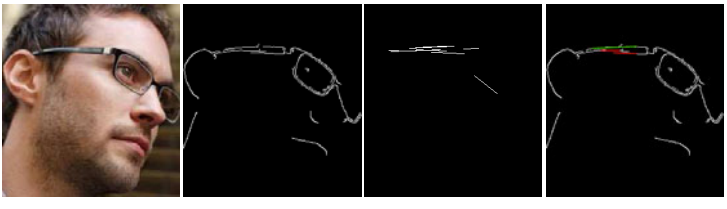


Fig. 5. Wide-Angle Eyeglass Leg Detection in Action from Original Image, Edge map, Line Candidates, and after Classification Process

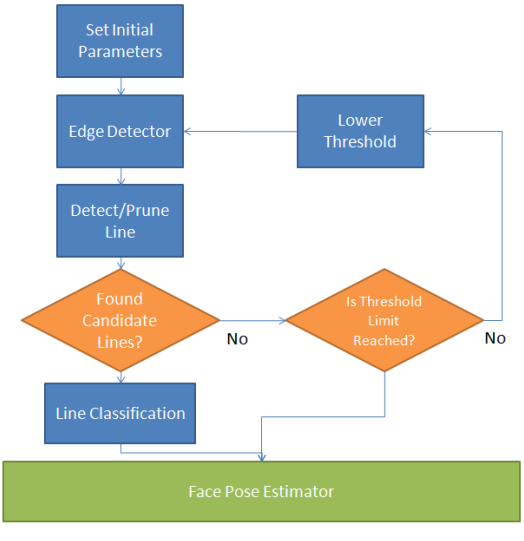


Fig. 6. Eyeglass Leg Detection Flowchart

Once the candidate lines are pruned, the remaining candidate lines are examined and classified. Lines that are in close vicinity with similar slopes are then grouped together as they are likely to be formed as the top and bottom segments of the eyeglass leg and reformed as super candidate lines which is illustrated in Fig. 5. However, it is a probability that there would be no candidate lines returned by this process.

In this case, it implies that the person is likely to be posing at a small angle in relationship with the camera, and due to that, no candidate lines were found in this process is possible. Once this step is done, then the super candidate line is then sent to Face Pose Estimation process.

4.4 Region Classification

The Region Classification process takes the ROI image of the subject head and generates different region maps that would be used in classification process. The first region is the face region map, which is created from using the skin color detector to create a skin map of the face. From the skin map, the contour of face is extracted via contour tracing. The largest contour is considered as the face region and other information could be extracted.

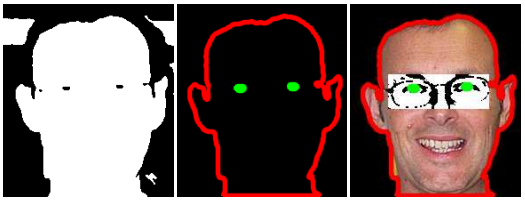


Fig. 7. Finding the Skin Region in the Region Classification Process. Skin Classification, Skin Contour classification, finding the Eyeglass Region using Eye Position and Face Region transpose on Adaptive Threshold of Face run Under Median Filter.

Another useful region is the eyeglass region which could be examined if the eyeglass legs are not detected which means the subject is likely in a frontal profile. Our approach is to utilize an eye detector algorithm that is based on skin color and intensity information [6][19]. Once the eye are found, the next step is to create a binary map from adaptive threshold of the ROI image that has a median filter run. Based on the eye position, and the binary map, the rough estimate of the extent of the eyeglass region can be calculated based on sampling of pixel density and edge information [10] [11] from the eye region outwards. Details of the process are displayed in Fig. 7.

4.5 Face Pose Classification

Once candidate lines have been detected from the previous process, the Face Pose Estimation process will perform a number of additional processing. The first step is to prune candidate lines that are unlikely to be formed by the eyeglass frame. Using the eyeglass region detected, candidate lines that are formed totally inside the region are pruned. The next step is to select the dominant line from the list of candidate lines that is likely to be formed from the eyeglass leg.

The next step is to classify which of the discrete pose the line suggests. Using the lines that are detected from the image and mapping them to the possible discrete poses, the angle and slope interval list is displayed at Table 1. In special cases, it is possible that the system would return with no candidate lines after the prune. In this case, then the person is likely to be at a frontal pose.

Table 1. Discrete Pose Classification Based on Angle and Slope Interval

Discrete Pose	Angle Interval	Slope Interval
Right	(-22.5°, 22.5°]	(-0.414,0.414]
TopRight	(22.5°, 67.5°]	(0.414,2.414]
Top	(67.5°,112.5°]	(2.414,-2.414]
TopLeft	(112.5°,157.5°]	(-2.414,-0.414]
Left	(157.5°,202.5°]	(-0.414,0.414]
BottomLeft	(202.5°,247.5°]	(0.414,2.414]
Bottom	(247.5°,292.5°]	(2.414,-2.414]
BottomRight	(292.5°,337.5°]	(-2.414,-0.414]
Straight	*(Special Case)*	

Due to the symmetric nature, additional processing has to be done to figure which of the symmetric direction the eyeglass leg pattern suggests. The first step is to figure the position of the eye. Our system uses a modified eye detector based on skin color and edge classification that is used to search along the interval of the dominant leg. Once the eye is detected, the direction can be figured based on the position of the eye and the general angle/slope of the leg. However there exist certain special cases in which the eyeglass legs may obscure the eyes or the eyes may not be detected properly. In this case additional processing has to be done.

To deal with that situation, the system samples an area for skin around the first and last point of the dominant line to figure which area has is likely to be the facing direction. As the ear areas typically contain less skin than the area close to the front, it is possible to detect which orientation the person is facing. By comparing between $SkinTotal_0$ and $SkinTotal_1$ from equation (1) and (2) respectively, it is possible to figure where the lens position is and deduce the orientation of the eyeglass on the subject.

$$SkinTotal_0 = \sum\sum Skin(x_0+i, y_0+j) \; . \tag{1}$$

$$SkinTotal_1 = \sum\sum Skin(x_1+i, y_1+j) \; . \tag{2}$$

Where $-0.1(image_{width}) \leq i \leq 0.1(image_{width})$, and $-0.1(image_{height}) \leq j \leq 0.1(image_{height})$

5 Results and Discussions

To test the proposed system, a set of images with faces are prepared and manually labeled as one of the 9 discrete poses. The images in the set are of high variety consisting of people in different facing poses, people that are wearing glasses, people from different ethical background, and taken from different cameras, lighting, and environmental conditions. A total of 497 pictures were chosen, consisting of 137 pictures from the Georgia Tech [5] Face Database, 127 from the Internet, 102 collected by the authors, 67 from the Face Tracer [13] database, and 30 pictures from the BioID [1] database, and 7 pictures from the Frontal Face [4] database. The results of the system are displayed in Table 2.

Table 2. Accuracy of Proposed System broken down by Discrete Pose

Discrete Pose	Total Pictures	Accuracy
Right	57	94.74%
TopRight	29	86.21%
Top	11	63.64%
TopLeft	21	90.48%
Left	64	93.75%
BottomLeft	33	93.94%
Bottom	13	69.23%
BottomRight	31	87.10%
Straight	189	93.65%
Total	403	91.29%

The results of the proposed system are rather promising. The system shows good accuracy on the detection of the face pose for many of the poses especially when the subject are facing left and right at high angles. Though the system works well, there are a number of issues that causes the detection rates to drop other than images. Other than usual issues such as bad contrast and lighting, the major contributing factor to the errors in the classification of the face poses is that certain hair style can cover the majority of the eyeglass legs as illustrated in Fig. 8. These hairstyles can potentially cause the eyeglass leg detection system to miss the detection of the eyeglass leg and classify the image in the wrong classification.

Fig. 8. Samples of Hair Styles and their Associated Edgemap that makes the Eyeglass Leg Detector unable to detect the Eyeglass legs

The major challenge of the proposed system is when the subject is viewing in the top or bottom pose. In this pose, there are two factors that cause issues with the system which is shown in Fig. 9. The first issue is that the eyeglass frame at this pose may not be interposed with the face. This causes issues in the system. The first issue is when the system may not adequate select an ROI region that encapsulates the glass region. The second issue is that the detection of the eyeglass is more difficult to detect when it is not interposed directly with the skin as factors such as hair or the background around the eyeglass. The last issue is that once the eyeglass leg has been detected, the orientation classification may wrongly classify due to the different nature of the eyeglass frame and its relationship with the face. Due to the combination of the 3 issues, the detection rates for poses at the top and bottom poses are lower than expected.

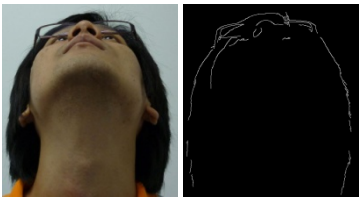


Fig. 9. Example of Top Pose in which the subject hair hides the eyeglass legs which leads to low eyeglass edge detection rates causing the system to incorrectly classify the pose

6 Conclusion and Future Work

This paper proposes a geometric method of estimating the head pose of a human subject that is wearing eyeglasses by using the novel feature of eyeglasses legs which is promising based on preliminary results. Based on the results, the system can estimate the head pose of a subject at a good accuracy rate, and works well with subjects having a high yaw angle in relationship with the camera.

Though the results are promising, there are some cases in which further studies need to be explored. For example, when a subject with long hair is facing in significant yaw angles, the hair has a tendency to obscure the eyeglass legs which confuses the proposed system. In this scenario, it is possible to fix such scenario by using a hybrid approach such as hair detection [23] to account for such scenario. Another case that needs further exploration is when the subject is at a high pitch angle such as looking straight up or down which requires further work. The exploration of more accurate eyeglass detection algorithms and the usage of the face skin profile can help increase the accuracy of the classification of the face pose. Another possible improvement for the proposed system is the implementation of the system to account for roll angles on the subject and make necessary changes. Since this system is using a geometric approach, it is possible to use geometric calculations to modify the algorithms proposed to account for the potential roll in the subject picture. One area to do additional work is to explore the area of perspective and how it affects the eyeglass leg patterns at different viewing angles in a roll, pitch, yaw perspective in more details.

Expanding the experiment by utilizing more images to building a larger image database is one area to improve on. The expanded database should be built with a fairer distribution of images poses, opposed as the present database that is skewed due to the rarity of available images that belong to certain discrete pose, could be done in the future.

References

1. BioID Face Database, `http://www.bioid.com/support/downloads/software/bioid-face-database.html`
2. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Active Shape Models – their training and applications. Computer Vision and Image Understanding 61(1), 38–59 (1995)
3. Cootes, T., Edwards, G., Taylor, C.: Active Appearance Models. IEEE Trans. Pattern Analysis and Machine Intelligence 23(6), 681–685 (2001)

4. Frontal Face Database,
http://www.vision.caltech.edu/Image_Datasets/faces/faces.tar
5. Georgia Tech Face Database,
http://www.anefian.com/research/gt_db.zip
6. Hassaballah, M., Ido, S.: Eye Detection Using Intensity and Appearance Information. In: MVA 2009 IAPR Conference on Machine Vision Applications, Tokyo, pp. 346–349 (2009)
7. Hammoud, R.I., Malawey, P.V.: Eyeglass Detection Method. U.S. Patent 7 370 970B2 (May 13, 2008)
8. Horprasert, T., Yacoob, Y., Davis, L.: Computing 3-d head orientation from a monoular image sequence. In: Int'l. Conf. Automatic Face and Gesture Recognition, pp. 242–247 (1996)
9. Huang, J., Shao, X., Wechsler, H.: Face pose discrimination using support vector machines (SVM). In: Int'l, Conf. Pattern Recognition, pp. 154–156 (1998)
10. Jiang, X., Binkert, M., Achermann, B., Bunke, H.: Towards Detection of Glasses in Facial Images. In: 141th International Conference on Pattern Recognition, pp. 1071–1073 (1998)
11. Jing, Z., Mariani, R.: Glasses Detection and Extraction by Deformable Contour. In: 15th International Conference on Pattern Recognition (ICPR 2000), vol. 2, pp. 29–33 (2000)
12. Kovac, J., Peer, P., Solina, F.: Human skin color clustering for face detection. In: EUROCON 2003, vol. 2, pp. 144–148 (2003)
13. Kumar, N., Belhumeur, P.N., Nayar, S.K.: FaceTracer: A Search Engine for Large Collections of Images with Faces. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 340–353. Springer, Heidelberg (2008)
14. Moon, H., Miller, M.: Estimating facial pose from a sparse representation. In: Int'l Conf. Image Processing, pp. 75–78 (2004)
15. Murphy-Chutorian, E., Trivedi, M.M.: Head Pose Estimation in Computer Vision: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(4), 607–626 (2009)
16. Nikolaidis, A., Pitas, I.: Facial Feature Extraction and Pose Determination. J. Pattern Recognition 33(11), 1783–1791 (2000)
17. Niyogi, S., Freeman, W.: Example-based head tracking. In: Int'l Conf. Automatic Face and Gesture Recognition, pp. 374–378 (1996)
18. Open Computer Vision Library,
http://sourceforge.net/projects/opencvlibrary/
19. Peng, K., Chen, L., Ruan, S., Kukharev, G.: A Robust Algorithm for Eye Detection on Gray Intensity Face without Spectacles. Journal of Computer Science & Technology (JCS&T) 5(3), 127–132 (2005)
20. Sherrah, J., Gong, S., Ong, E.J.: Face distribution in Similarity space under varying head pose. J: Image and Vision Computing 19(12), 807–819 (2001)
21. Umbaugh, S.E.: Computer Imaging, Digital Image Analysis and Processing. Taylor & Francis, Boca Ranton (2005)
22. Vezhnevets, V., Sazonov, V., Andreeva, A.: A survey on pixel-based skin color detection techniques. In: GRAPHICO 2003, pp. 85–92 (2003)
23. Yacoob, Y., Davis, L.: Detection, Analysis and Matching of Hair. In: Tenth IEEE International Conference on Computer Vision (ICCV 2005), vol. 1, pp. 741–748 (2005)
24. Wang, J.G., Sung, E.: EM enchancement of 3D head pose stimulated by point at infinity. Image and Vision Computing 25(12), 1864–1874 (2007)
25. Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-time combined 2D+3D active appearance models. In: IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 535–542 (2004)
26. Zhao, G., Chen, L., Song, J., Chen, G.: Large head movement tracking using SIFT-based registration. In: Proc. Int'l Conf. Multimedia, pp. 807–810 (2007)

Feature Selection for Neural Network Based Stock Prediction

Prompong Sugunnasil and Samerkae Somhom

Department of Computer Science, Faculty of Science,
ChiangMai University, ChiangMai 50200, Thailand
p.sugunnasil@chiangmai.ac.th,
samerkae@chiangmai.ac.th

Abstract. We propose a new methodology of feature selection for stock movement prediction. The methodology is based upon finding those features which minimize the correlation relation function. We first produce all the combination of feature and evaluate each of them by using our evaluate function. We search through the generated set with hill climbing approach. The self-organizing map based stock prediction model is utilized as the prediction method. We conduct the experiment on data sets of the Microsoft Corporation, General Electric Co. and Ford Motor Co. The results show that our feature selection method can improve the efficiency of the neural network based stock prediction.

Keywords: Feature Selection, Stock Prediction, Correlation, Neural Network.

1 Introduction

The stock prediction has always been a historical popular subject for financial computing. Many approaches have been applied to predict stock price and movement. They include artificial neural networks [17][3][22][10], Bayesian belief network [20], genetic algorithm [8][9][12], classifier system [13], fuzzy set [18][4], associative rule [14] and hybrid method that combines a few approaches [1][11][5][2]. Every approach is concerned with large datasets which consume a lot of computation time; moreover, the large datasets sometimes deteriorate the predictions accuracy. The feature selection can be utilized to reduce the number of input by evaluating and selecting the suitable choice of data features. The effective data feature is not only reduce the dimension of the input space, but also can improve the quality and performance such as redundancy reduction, noise elimination, improvement of processing speed and even facilitation of data modeling.

Researchers have developed a large number of feature selection techniques for various purposes. According to [24], there are three categories of feature selection method: *wrapper*, *filter* and *embedded method*.

Wrapper methods utilize the learning machine of interest to score the subset of feature. [19] proposed a feature selection method for SVM. The method is based

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 137–146, 2010.

upon finding the feature(s) minimizing the bounds on leave-on-out error. The research shows that the performance is superior to some standard feature selection, Pearson correlation coefficient, the Fisher criterion score, and the Kolmogorov-Smirnov test. Some of the researcher consider wrapper method as one the brute force method which require a lot of computation time but we can apply search method to reduce the computation time.

Filter methods is a tool used to select the subset of feature based on the properties of data. The filter methods function as the pre-processing step. [15] proposed a feature selection method based on correlation. The research describes a correlation based heuristic method to score the feature(s) and evaluate them with C4.5 algorithm, naive Bayes classifier,and instance based learner. The result of the study show that the proposed method can improve the accuracy of the machine learning algorithm and also perform many time faster than wrapper method.

Embedded methods perform the feature selection within the training process and are usually specific to given learning machines. The Optimal Brain Damage (OBD)[23] is a method to reduce the size of a learning network for neural network. The OBD selectively deletes the weight and results in a neural network work well or better.

The self-organizing map (SOM) has been widely used as a prediction with a great deal of success. The SOM is an unsupervised neural network which frequently used for classification and clustering. SOM can be used to cluster of data without any priori knowledge of number of cluster; moreover, the result of SOM as clustering method is better than other hierarchical clustering method[6]. In [25], the authors also suggest that the SOM seems to be faster than other traditional methods without diminishing the quality.

Our contribution of this paper is to propose a feature selection method for SOM based stock prediction. The feature selection method results from the study of technical indexes of stock and SOM. In the next section, we formally define the feature selection problem. Section 3 gives an overview over SOM-based prediction method. In Section 4, we introduce our feature selection method. The experiment and the result is presented Section 5. In the last section, we conclude this paper and discuss the future work.

2 The Feature Selection Problem

The feature selection problem is an optimization problem that selects subset of features. The selected subset should produce an acceptable rate of error. We follow the problem description in [19]. Given the prediction function $f(X)$ where X denotes an set of data with size m and Y denotes an set of target output with the same size as X, the feature selecting parameter α with size n that smaller than m and has value of zero or one $\alpha \in \{0,1\}^n, \lambda(X,Y)$is loss function, we wish to find the parameter λ that give the minimum error value. Let the feature selection function $\tau(X,\alpha)$ be

$$min\left(\int \lambda\left(Y, f\left(X \times \alpha\right)\right) d\lambda\left(Y, f\left(X \times \alpha\right)\right)\right) \tag{1}$$

3 The SOM Based Stock Prediction Model

In this paper, we apply the extension of the system presented in [22] as prediction method by changing the input from trading data to technical indexes. We try to model the behavior of stock in the same period because there are various factors that influence stock prices and each factor has different effects at a time. If we could model the behavior of stock, we should predict the movement of stock[16].

We first normalize the technical indexes into percent of change form using equation (2). And, we construct the target output from equation (3).

$$x_{new}^{i}(t) = \frac{x_{old}^{i}(t) - min(i)}{max(i) - min(i)} \tag{2}$$

where $x_{old}^{i}(t)$ is the old data of dimension i at the time t, $min(i)$ is minimum value of dimension i, $max(i)$ is maximum value of dimension i, $x_{new}^{i}(t)$ is the new data of dimension i at the time t.

$$A(t+1) = \frac{f(t+1) - f(t)}{f(t)} \times 100 \tag{3}$$

where $A(t)$ is the target output at time t, $f(t)$ is closing price data at time t

After we obtain the input and the target output, we cluster the target output into 3 groups which represent the periods of holding, selling or buying. We group the holding period and selling period together in order to distinguish between the interesting group, the price will increase and un-interesting group, the price will increase.

After we have the groups, we utilize the SOM as prediction tool by learning the characteristics of each group. When we predict the stock, we first use historical data from a point in the history to the present time to construct the model and we use the closing price of the present time as the testing data for the future. When the model guides that the way of the stock price will increase, we should buy. And when the model guides that the way of the stock price will decrease, we should sell or hold.

4 The Proposed Feature Selection Method

In this section, we describe our feature selection method. Like other feature selection approach, we divide our process into 2 stages: Search and Evaluation. The objective of search step is to search through the list of features for the best subset. And, the objective of evaluation step is to evaluate the choice of feature.

4.1 Assumption

Our assumptions are based on association between each feature in the same subset, association of feature and the target output and the characteristic of the feature.

According to the neural network theory, the special characteristic of neural networks' input is that input variables should not be much correlated, because the correlated input variables may degrade the prediction performance by interacting each other as well as other elements and producing the biased[21]. In other words, the neural network may get confused with the contributed information[7].

The relation among inputs and the relation between input and target output or input can be quantified by using the standard Pearson correlation which is described by

$$corr\,(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{4}$$

where X denotes the input variables, Y denotes the target output, x_i is the input variable at time i, y_i is the target output at i, n denotes the size of X, $corr(X,Y)$ is the correlation function of X and Y.

The characteristics of the feature also have effect on prediction's performance. The redundant information from the input variable itself also deteriorates the prediction performance. As is well known, the autocorrelation is the popular measurement for the correlation inside each the time series data. The autocorrelation is described by

$$autocorr\,(X,k) = \frac{\sum_{i=k+1}^{n}(x_i-\bar{x})\,(x_{i-k}-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{5}$$

where X denotes the input variable, k denotes the lag periods, $autocorr(X,k)$ is the autocorrelation function of X at lag period of size k.

4.2 Evaluation Function

From the assumption mentioned in previous section, we can conclude our input selection into a single objective function :

$$min\left(\left|\frac{\sum_{i=1;j=1;i\neq j}^{\binom{m}{2}} cor\,(x_i,x_j)\sum_{i=1}^{m} autocorr\,(x_i)}{\sum_{i=1}^{m} cor\,(x_i,y)}\right|\right) \tag{6}$$

where m denotes the size of input variables, x_i denotes the i^{th} input variable, y denotes the target output.

We only focus on the size of relation. So we apply absolute to get rid of the direction.

4.3 Search

Since the number of generated input subsets is exponential in size, the sequential search is usually intractable. So, we apply heuristic approach in order to

reduce the computation time[15]. The Hill climbing search algorithm is shown as follows.

Hill climbing search algorithm

```
program Hill climbing search
   currentValue :=  firstValue;
   loop
      nextValue := NEXT(currentV alue);
      newValue := NULL;
      eval := -1;
      for all t in nextValue do
         if EVALUATION(x) > eval then
            newValue := x;
            eval := EVALUATION(x);
         end if
      end for
      if eval <= EVALUATION(currenValue) then
         return currentValue;
      end if
   end loop;
end.
```

5 The Experiments and Results

We use 46 technical indexes of Microsoft Corporation, General Electric Co. and Ford Motor Co. The example of technical index is shown in Table 1. We obtain the trading data from YAHOO[1]. The trading data were reported daily from 14 September 2005 to 13 February 2008 total 609 records. We used learning rate parameter of 0.01 and initial weights of 0.01. We used different size of time frame in order to determine the performance of our feature selection method in different environment.

The table 2 to table 4 show the performances of the feature selection methods on the datasets. In these experiments, we compare the accuracy of prediction without feature selection, the accuracy of prediction with our feature selection method by using two features and the accuracy of prediction with our feature selection method by using three features at different time frame.

We also conducted experiment on the influential factor of selected feature by factorial design at various confidential level (90%, 95%, 99%, 99.9%, 100%). We provide example of experiment of three datasets from 17 March 2008 to 11 June 2008. In this experiment, we used time frame of 60 day and selected three features. The result is shown in table 6 to table 8. The meaning of significant is shown in table 5.

[1] http://quote.yahoo.com

Table 1. Example of technical index

Technical index	Description
Acceleration	Measure the acceleration of the current driving force.
Accumulation/Distribution	Calculate from the changes in price and the volume.
Aroon Indicator	Determine whether a stock is trending or not and how strong the trend is.
Bollinger Bands	plotted a certain number of standard deviations away from the moving average.
Commodity Channel Index	Identify cyclical turns in commodities.

Table 2. Average accuracy of General Electric dataset prediction

Time Frame(day)	Accuracy(%)		
	Whole	Our method	
		2 features	3 features
7	68.61	70.32	71.13
15	65.35	65.25	68.59
30	60.51	63.64	68.18
45	57.27	63.74	64.65
60	59.27	65.63	69.06
75	55.90	67.97	67.78

Table 3. Average accuracy of Microsoft Corporation dataset prediction

Time Frame(day)	Accuracy(%)		
	Whole	Our method	
		2 feature	3 feature
7	74.25	70.32	74.54
15	68.08	65.35	65.56
30	66.57	60.30	63.03
45	65.05	60.00	60.60
60	64.79	62.29	63.13
75	66.78	60.00	62.77

From the results of experiments, we make the following observations: 1) Our feature selection method can improve the efficiency for the General Electronic dataset. The results show that the accuracy of the prediction with our feature selection method can achieve the same or even better performance when compares to the prediction without any feature selection method. With more features selected, the accuracy can be even better. 2) Our feature selection method with slightly deterioration of accuracy in Microsoft Corporation and Ford Motor Co.

Table 4. Average accuracy of Ford Motor Co. dataset prediction

Time Frame(day)	Accuracy(%)		
	Whole	Our method	
		2 feature	3 feature
7	73.95	70.52	73.54
15	71.71	65.15	65.45
30	69.09	60.20	64.84
45	67.67	61.91	64.54
60	65.94	60.42	61.98
75	65.64	65.64	65.13

Table 5. Percent of confidential

Symbol	Percent of confidential(%)
'***'	100
'**'	99.9
'*'	99
'.'	95

Table 6. Significant of selected feature from General Electric dataset

Feature	Mean Sq.	F value	Pr(>F)	Signif.
1[a]	138.738	65.9780	8.150e-11	***
2[b]	68.557	32.6027	5.508e-07	***
3[c]	15.556	7.3978	0.0088580	**
1:2	27.496	13.0758	0.0006758	***
2:3	3.386	1.6101	0.2101309	
1:3	5.972	2.8398	0.0979491	
1:2:3	1.351	0.6425	0.4264682	

[a] Average Change
[b] Breadth Adv./Decl.
[c] Range Diff.

Table 7. Significant of selected feature from Microsoft Corporation dataset

Feature	Mean Sq.	F value	Pr(>F)	Signif.
1[a]	3.5929	12.1223	0.001018	**
2[b]	21.1714	71.4319	2.477e-11	***
3[c]	1.6752	5.6520	0.02114	*
1:2	1.4280	4.8181	0.032650	*
2:3	1.9079	6.4372	0.014214	*
1:3	0.0994	0.3353	0.565070	
1:2:3	0.0107	0.0361	0.85	

[a] Advance Decline Line
[b] Chaikin A/D Osc.
[c] PAIN

Table 8. Significant of selected feature from Ford Motor Co. dataset

Feature	Mean Sq.	F value	Pr(>F)	Signif.
1[a]	5.6975	12.6156	0.0008229	***
2[b]	1.0090	2.2341	0.1410437	
3[c]	0.0010	0.0022	0.9623607	
1:2	5.1978	11.5090	0.0013310	**
2:3	4.9003	10.8504	0.0017817	**
1:3	0.2172	0.4810	0.4910584	
1:2:3	12.434	27.5326	2884e-06	***

[a] Average Positive Change
[b] Breadth Adv./Decl.
[c] Chaikin Money Flow

dataset, still the prediction result can be improved by selecting more features. 3) For each selection, our feature selection method is able to select the feature with confidential at least 99%. To figure out the real reason, we need more experiments.

6 Conclusion

This paper presents a feature selection method for neural network based stock prediction that utilizes correlation to measure the appropriateness of the input variables. In this paper, we formalized the feature selection as the optimization problem. We try to maximize the correlation between input and target output and to minimize the correlation among the selected input variable and autocorrelation of input variable. The advantage of our method is data-driven that the prior knowledge is not necessary. Results show that the method can choose subset of feature with significant at 90% confidence and can maintain the prediction's accuracy by using smaller size of dataset.

The future work will attempt to understand why the proposed method works efficiently in some case than other. Comparing the proposed method with other feature selection method and with different machine learning algorithm will be explored.

References

1. Afolabi, M.O., Olude, O.: Predicting stock prices using a hybrid kohonen self organizing map (som). In: Proc. of the 40th Annual Hawaii International Conference on System Sciences, p. 48. IEEE Press, New York (2007)
2. Ao, S.I.: Automating stock prediction with neural network and evolutionary computation. In: Liu, J., Cheung, Y., Yin, H. (eds.) IDEAL 2003. LNCS, vol. 2690, pp. 203–210. Springer, Heidelberg (2003)
3. Kimoto, T., Asakawa, K., Yoda, M., Takeoka, M.: Stock market prediction system with modular neural networks. In: IJCNN: International Joint Conference on Neural Networks, pp. 1–6. IEEE Press, New York (1990)
4. Chang, P.C., Liu, C.H.: A tsk type fuzzy rule based system for stock price prediction. Expert Syst. Appl. 34, 135–144 (2008)
5. Fu, J., Lum, K.S., Nguyen, M.N., Shi, J.: Stock prediction using fcmac-byy. In: Liu, D., Fei, S., Hou, Z., Zhang, H., Sun, C. (eds.) ISNN 2007. LNCS, vol. 4492, pp. 346–351. Springer, Heidelberg (2007)
6. Lin, G.F., Wanga, C.M.: Performing cluster analysis and discrimination analysis of hydrological factors in one step. Adv. Water Resour. 29, 1573–1585 (2006)
7. Huang, W., Wang, S., Yu, L., Bao, Y., Wang, L.: A new computational method of input selection for stock market forecasting with neural networks. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006. LNCS, vol. 3994, pp. 308–315. Springer, Heidelberg (2006)
8. Kaboudan, M.A.: Genetic programming prediction of stock prices. Comput. Econ. 16, 207–236 (2000)
9. Kim, K.j., Han, I.: Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. Expert Systems with Applications 19, 125–132 (2000)
10. Kwon, Y.K., Choi, S.S., Moon, B.R.: Stock prediction based on financial correlation. In: Proc. of Genetic and Evolutionary Computation Conference, pp. 2061–2066. ACM, Washington (2005)
11. Kwon, Y.K., Moon, B.R.: Daily stock prediction using neuro-genetic hybrids. In: Proc. of Genetic and Evolutionary Computation Conference, pp. 2203–2214. ACM, Washington (2003)
12. Kwon, Y.K., Moon, B.R.: Evolutionary ensemble for stock prediction. In: Proc. of Genetic and Evolutionary Computation Conference, pp. 1102–1113. ACM, Washington (2004)
13. Liao, P.Y., Chen, J.S.: Dynamic trading strategy learning model using learning classifier systems. In: Proc. of the 2001 IEEE Congress on Evolutionary Computation, pp. 783–789. IEEE Press, New York (2001)
14. Lu, H., Han, J., Feng, L.: Stock movement prediction and n-dimensional intertransaction association rule. In: Proc. of the ACM SIGMOD Workshop on IB Issues on Data Mining and Knowledge Discovery, pp. 1–7. ACM, Washington (1998)

15. Hall, M.A., Smith, L.A.: Feature subset selection: a correlation based filter approach. In: Proc. of the 1997 International Conference on Neural Information Processing and Intelligent Information Systems, pp. 855–858. Springer, Heidelberg (1997)
16. Sugunsil, P., Somhom, S.: Short term stock prediction using som. In: Proc. of United Information Systems Conference 2009. LNBIP, vol. 20, pp. 262–267. Springer, Heidelberg (2009)
17. Tino, P., Schittenkopf, C., Dorffner, G.: Financial Volatility Trading using Recurrent Neural Networks. In: Proc. of IEEE Transactions on Neural Networks, pp. 865–874. IEEE Press, New York (2001)
18. Wang, Y.F.: Predicting stock price using fuzzy grey prediction system. Expert Syst. Appl. 22, 33–38 (2002)
19. Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., Vapnik, V.: Feature selection for support vector machines. In: Advances in Neural Information Processing Systems, vol. 13, pp. 668–674. MIT Press, Cambridge (2001)
20. Wolfe, K.R.: Turning point identification and bayesian forecasting of a volatile time series. Comput. Ind. Eng. 15, 378–386 (1988)
21. Zhang, G.P.: Neural Networks in Business Forecasting. Information Resources Press, VA (2003)
22. Zorin, A.V.: Stock Price Prediction: Kohonen Versus Backpropagation. In: Proceeding of International Conference on Modelling and Simulation of Business Systems, pp. 115–119. IEEE Press, New York (2003)
23. LeCun, Y., Denker, J.S., Solla, S.A.: Optimal brain damage. In: Advances in Neural Information Processing Systems, pp. 598–605. Morgan Kaufmann, San Francisco (1989)
24. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 3, 1157–1182 (2003)
25. Lendasse, A., Verleysen, M., Bodt, E.: Forecasting Time-Series by Kohonen Classification. In: Proceedings of European Symposium on Artificial Neural Networks, pp. 221–226. D-Facto public., Bruges (1998)

Design Concept for Garbage Bin with Situation Awareness Feature

Montri Supattatham and Borworn Papasratorn

School of Information Technology
King Mongkut's University of Technology Thonburi
Thailand
{montri,borworn}@sit.kmutt.ac.th

Abstract. Many measures to prevent wide-spread of communicable diseases depends on embedded IT into objects found in public places. This makes it possible to have objects with awareness on surrounding environment, or having situation awareness. This paper presents design concept to add situation awareness features to automatic garbage bin. There are three design levels for including situation awareness features with garbage bin. From awareness goals, required features are identified. Perception, comprehension, and projection are then aligned with the required features, in order to have desired awareness. Automatic garbage bin is implemented using design specification from the proposed design concept. Result from convenience sampling survey reveals that users are satisfied with the implemented garbage bin.

Keywords: Situation Awareness, automatic garbage bin, embedded system, design concept.

1 Motivation

One of threat facing human society at present is emerging communicable diseases. Outbreak of SAR and H5N1 flu recently have created number of measures to prevent wide- spread of diseases. Transmission of an infectious disease may occur through diverse pathways, especially direct physical contact with infected individuals. These infecting agents may also be transmitted through liquids, food, and body fluids. Contaminated objects from liquids and body fluids exist almost every-where in public place. Lots of measures to prevent direct contact with infected persons and contaminated objects are widely implemented. Some of the measures depend on IT systems, e.g. scanning of human temperature upon entry into public places. Embedded systems also have roles in reducing presence of pathogenic microbial agents in public place. One example is automatic alcohol dispenser. One place we can find the contaminated objects is garbage bin, which are existed every-where in public place. There are numbers of effort to reduce human contact upon using garbage bin [1,2]. Automatic garbage bins help reduce inflection from direct contact with contaminated objects from body fluids, as well as making it becomes convenience.

With advance in IT, there are many research related to augmentation of objects in public places with IT to enable objects to obtain general context information within

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 147–153, 2010.

their environment. This will create objects that enhance human convenience in public place. Objects are integrated into specific computer supported tasks, such as smart coffee cups[3], augmented trash can [4], camera sensor network [5],etc. . This advance makes it possible to have objects with awareness on surrounding environment, or having situation awareness. Situation awareness has many definition [6,7]. According to Endley [8], Situation awareness is defined as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" One of the key benefits of situation awareness is that it informs what data is needed and how to combine to make situation understood. To enable objects to understand situation, IT have to be embedded into objects. Specification and design of embedded systems demands new approaches beyond traditional hardware design concept. Development of the Unified Modeling Language (UML) proposes for the new design concept, however it still unfamiliar for hardware designer [9].

This paper presents design concept to add situation awareness to automatic garbage bin. The proposed design concept follows the layering process which is well understood among hardware designer. With situation awareness, the implemented garbage bin can function properly without human contact. This will enhance user's intention to use, which will eventually increase usage. The proposed garbage bin can therefore help reducing human contact with the contaminated objects, as well as making it easy to design and make.

2 Situation Awareness Design Concept

According to Endley [8], situation awareness consists of 3 levels: Perception of the elements in the environment, comprehension of the current situation, and projection of future status. To enable garbage bin to have awareness about surrounding situation, we associate 3 levels of design with desired features as shown in Fig. 1.

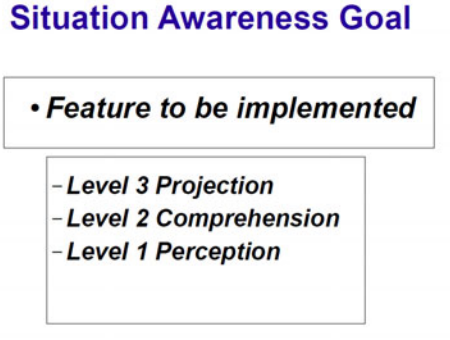


Fig. 1. Design Layer to add situation awareness to garbage bin

Design begins with setting awareness to be included in garbage bin. In our design, we specify 3 awareness for the proposed garbage bin: awareness of garbage, awareness of user, and awareness of environment. Once awareness goal is assigned, features to be included will be identified. We then align situation awareness levels to the

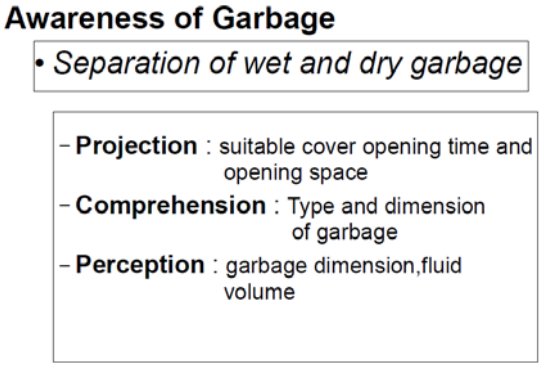


Fig. 2. Alignment of garbage awareness to three design levels

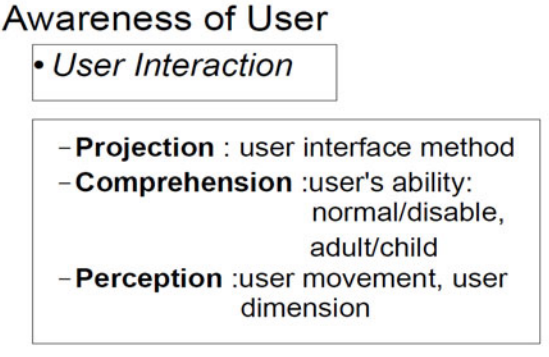


Fig. 3. Alignment of user awareness to three design levels

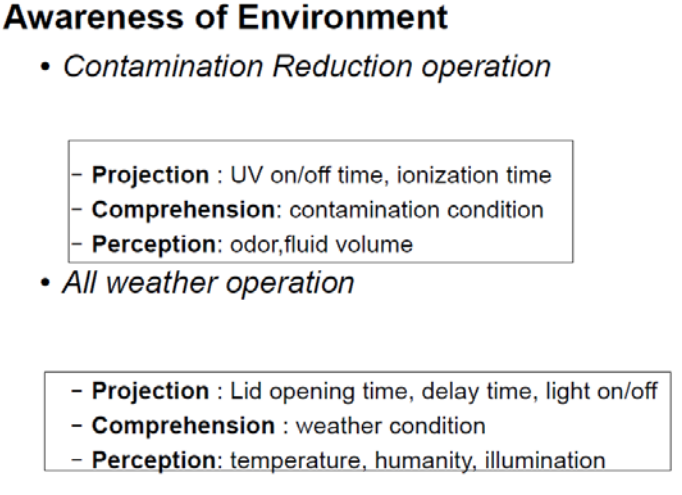


Fig. 4. Alignment of environmental awareness to three design levels

identified features. The first level is projection of awareness required to implement the features. The second level is to understand what situation is at present. The third level identifies required data to make situation being understood. Garbage bin can separate wet and dry garbage using awareness of garbage condition. Suitable user interface is selected based on user awareness. General operation and contamination reduction function according to awareness of environment. Alignment between each awareness goals and three design levels are shown in Fig. 2, 3, 4 respectively.

3 Design Specification of Garbage Bin with Situation Awareness

From the proposed design concept, we made alignment between awareness goals and features, some examples of alignment are shown in Fig. 5, 6, 7.

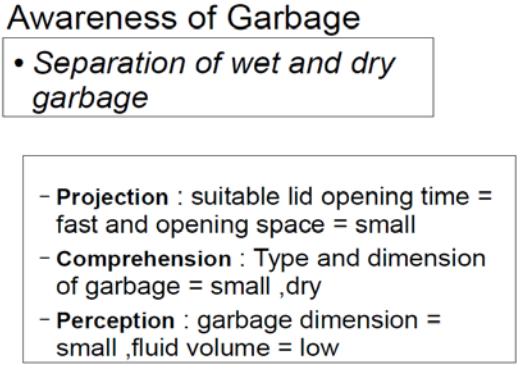


Fig. 5. Example of design level for garbage awareness

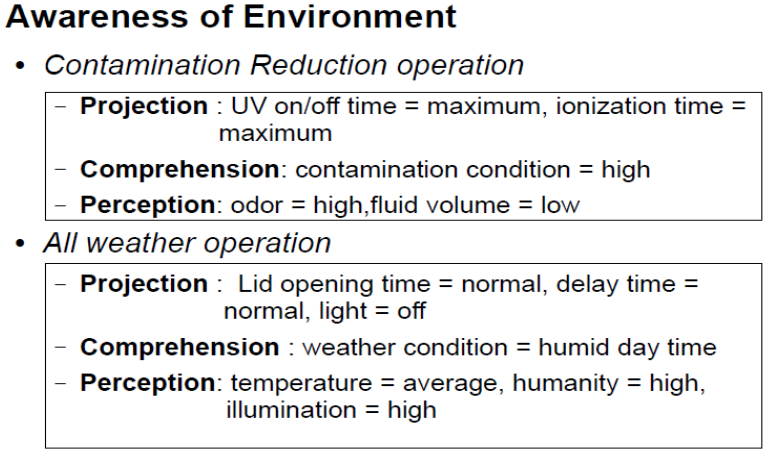


Fig. 6. Example of design level for user awareness

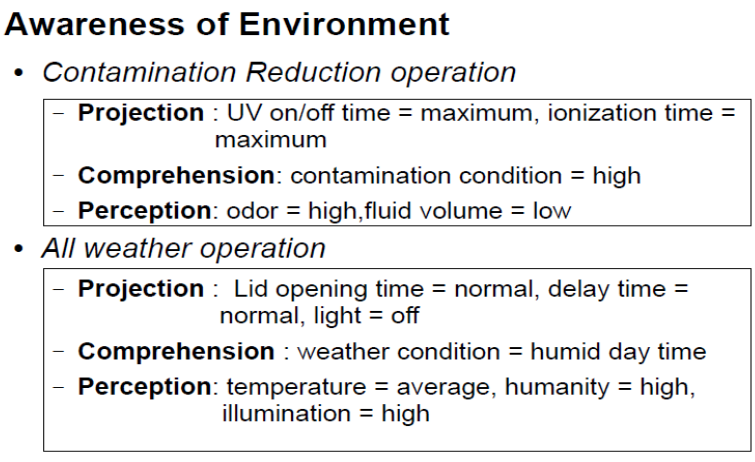


Fig. 7. Example of design level for environmental awareness

From the alignment, we have specification for automatic garbage bin as summarized in Table 1. The implemented garbage bin partially covers the proposed specification, as shown in Fig. 8. The garbage bin was located in canteen of the School of Information Technology, KMUTT to test user's intention to use and satisfaction. With population of 100, we made random convenience survey with sample size of 20. Four from five users have expressed intention to use the proposed garbage bin and three from five users satisfy with its features and operation during 6 months period.

Table 1. Design specification of the proposed garbage bin

Features	Specifications
Dimension	Covers 2X 100 Liters standard plastic tanks
Number of Tanks	2 (1 for wet, 1 for dry)
Material	Outside: Stainless steel, Inside: Plastic
Detection of User's intention to use	2 points movement sensors
Cover opening time and space	Fixed or automatic assigned according to awareness in garbage, user, and environment
Contamination reduction	UV and Ionization with on-off time dependent on awareness
User Interface	Automatic or Voice interaction

Fig. 8. Automatic Garbage bin with situation awareness features

4 Conclusion and Future Research

Reducing human contact with contaminated objects can help prevent spreading of communicable diseases. We propose design concept for automatic garbage bin with situation awareness features. Awareness of three related entities are identified and aligned with three design levels. The proposed concept makes design easy to understand and implement, when compared to other sophisticated design concepts. Although the proposed design concept covers variety of situations, further research is required for better understanding of various usage scenario. Design specification will be more complete once we can test our design specification with usage in various scenarios. This will reduce possibility of failure design and increase user's satisfaction on embedded objects for everyday usage.

References

1. Best Touchless Trash Can - iTouchless Automatic Stainless Steel 13 Gallon Trash Can (2010), http://hubpages.com/hub/Touchless-Trash-Can-iTouchless-Stainless-Steel-13-Gallon-Trash-Can
2. Daping, L.I., Weiping, L.I.U., Lichao, P.A.N., Daqiang, G.U.: The Design of a Secure Automatic Garbage Bin for Hospitals. Journal Information: World Sci-Tech R & D (6), 819–821 (2008)
3. Gellersen, H.-W., Beigl, M., Krull, H.: The MediaCup:Awareness Technology embedded in an Everyday Object. In: Gellersen, H.-W. (ed.) HUC 1999. LNCS, vol. 1707, pp. 308–310. Springer, Heidelberg (1999)
4. JetSam: An Urban Probe, Augmented Trash Can (2010), http://www.urban-atmospheres.net/UrbanProbes/Jetsam/artifact.htm
5. Shin, U., Kumar, R., Mohapatra, D., Ramachandran, U., Ammar, M.: ASAP: A Camera Sensor Network for Situation Awareness. In: Tovar, E., Tsigas, P., Fouchal, H. (eds.) OPODIS 2007. LNCS, vol. 4878, pp. 31–47. Springer, Heidelberg (2007)
6. Beringer, D.B., Hancock, P.A.: Exploring situational awareness: A review and the effects of stress on rectilinear normalisation. In: Proceedings of the Fifth International Symposium on Aviation Psychology, vol. 2, pp. 646–651 (1989)

7. Green, M., Odom, J.V., Yates, J.T.: Measuring situational awareness with the “Ideal Observer”. In: Proceedings of the International Conference on Experimental Analysis and Measurement of Situation Awareness. Embry-Riddle Aeronautical University Press (1995)
8. Endsley, M.R.: Situation awareness global assessment technique (SAGAT). In: Proceedings of the National Aerospace and Electronics Conference (NAECON), pp. 789–795. IEEE, New York (1988)
9. Martin, G.: UML for Embedded Systems Specification and Design: Motivation and Overview. In: Design Automation and Test in Europe Conference and Exhibition (DATE 2002), p. 0773 (2002)

Application of Cellular Automata in Symmetric Key Cryptography

Miroslaw Szaban[1], Jerzy Pawel Nowacki[2], Aldona Drabik[2], Franciszek Seredynski[2,3], and Pascal Bouvry[4]

[1] Institute of Computer Science, University of Podlasie,
3-go Maja 54, 08-110 Siedlce, Poland
mszaban@ap.siedlce.pl
[2] Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland
{nowacki,adrabik,sered}@pjwstk.edu.pl
[3] Institute of Computer Science, Polish Academy of Sciences,
Ordona 21, 01-237 Warsaw, Poland
sered@ipipan.waw.pl
[4] Faculty of Sciences Technology and Communication,
University of Luxembourg
6 rue Coudenhove Kalergi, L-1359 Luxembourg, Luxembourg
pascal.bouvry@uni.lu

Abstract. The main concern of this paper is cryptosystems with a symmetric key, in particular block ciphers. The most important components of block ciphers are substitution boxes (S-boxes). Developing methods of cryptanalysis make ciphers worked on classical S-boxes not safe enough. Therefore, we propose a methodology of generation cellular automata (CA)-based S-boxes with enhanced quality. We provide an exhaustive experimental analysis of the proposed CA-based S-boxes in terms of non-linearity, autocorrelation, balance and strict avalanche criterion. We show that proposed S-boxes have high cryptographic quality. The interesting feature of the proposed S-boxes is a dynamic flexible structure, fully functionally realized by CA, while the classical S-boxes are represented by fixed table structures.

Keywords: Cellular Automata, S-boxes, Block Cipher, Cryptography, Boolean Functions.

1 Introduction

This paper is dedicated to cryptosystems with a secret key (symmetric key), in particular block ciphers. Block ciphers operates on fixed-length groups of bit-termed blocks, with use of a fixed transformation [7]. The base components of the block ciphers are S-boxes (specially designed numerical tables), which serve as tools of nonlinear transformation of information in the cipher process.

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 154–163, 2010.

The $n \times k$ S-box (see [5]) is a function f: $B^n \rightarrow B^k$ which from each of n Boolean input values of B^n block consisting of n bits b_i ($i \leq n$) generates some k Boolean output values called B^k block consisting of k bits b_j ($j \leq k$ and $k \leq n$), what corresponds to the mapping bit strings $(b_1, b_2, \ldots, b_n) \rightarrow (b_1, b_2, \ldots, b_k)$.

One of well-known application of S-boxes is using them in DES as the 'heart' of this algorithm [5]. Each of eight wide known DES S-boxes *S1, ..., S8* is a table composed of 16 columns and 4 rows, and maps *6* input bits into *4* output bits. So, these eight functions collectively transform the *48*-bit input block into *32*-bit output block (see [5]).

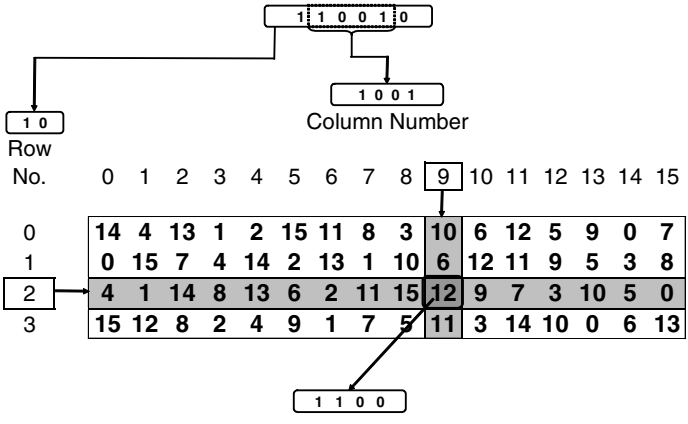


Fig. 1. Mapping the 6-bits into the 4-bits with use of S-box *S1* represented as a table in DES algorithm [5]

Let us consider the function *S1* represented in Fig. 1 by a specially designed table. Suppose that the input block of this function is the block B^6, e.g. *110010*. Two bits from B^6, the first and the last one (bits *10*) define row *2* of the *S1* block. Four middle bits *1001* define the column *9* of the *S1* block. Intersection of the column *9* and row *2* points in the table the number *12*, what corresponds to *1100*, and these bits are considered as the B^4 output block.

S-boxes are also used in modern symmetric key cryptography systems, e.g. in the new standard AES [6], successor of DES, and more other systems.

Let us note that the classical S-boxes, as described above, are fixed, not flexible structures requesting predefined sizes of memory. Therefore, it is hard to use them in new designed cryptographic algorithms, which request using dynamic S-boxes. The purpose of this study was to design flexible S-boxes, ready to use in cryptographic algorithms with dynamic S-boxes. It seems that CA are appropriate tools to design such S-boxes. CA are computationally universal (see e.g., [11], [2], [15]), what means that such Boolean functions can be realized. Furthermore, CA of a given size and with their rules (see Section 5) can potentially realize not one, but a number of S-box functions, what gives a possibility of designing much stronger cryptography systems. CA is a highly parallel system, easy in hardware implementation, what results in a high efficiency of CA-based cryptographic systems.

The papers are organized as follows. Section 2 describes the main cryptographic criteria to examine Boolean functions. Section 3 outlines the concept of CA. In Section 4 the idea of creating dynamic CA-based S-boxes is proposed. Section 5 presents analysis of CA in the context of tuning parameters for designing dynamic CA-based S-boxes. The analysis of cryptographic properties constructed dynamic CA-based S-boxes are provided in Section 6. The last section concludes the paper.

2 Cryptographic Criteria for Evaluation of Boolean Functions

The quality of S-boxes, also designed with the use of CA must be verified by required properties of S-boxes. The most important definitions and dependencies related to this issue are recalled below from cryptographic literature [1], [3], [4], [14].

A Boolean function $f: B^n \rightarrow B$, maps n binary inputs to a single binary output. The list of the size of 2^n of all possible outputs is the *truth table*. *Polarity* form of the *truth table* is denoted by $\hat{f}(x)$ and defined as $\hat{f}(x) = (-1)^{f(x)}$.

The non-linearity N_f of a Boolean function f is the minimal distance of the function f to the set of affine functions and is calculated as

$$N_f = \frac{1}{2}(2^n - WH_{\max}(f)) \cdot \tag{1}$$

Ciphers with high non-linearity (low WH_{max}) are known to be more difficult to cryptanalysis (more secure).

The next important property of ciphers is autocorrelation AC_f. Autocorrelation transform defines correlation between polar form $f(x)$ and its polar shifted version, $f(x \oplus s)$, where the operation $\oplus$ denotes XOR: bit-by-bit addition modulo 2. The absolute maximum value of any autocorrelation transform is defined as an autocorrelation and denoted by the equation:

$$AC_f = \max_{s \neq 0} \left| \sum_x \hat{f}(x) \hat{f}(x \oplus s) \right|, \tag{2}$$

where $s \in B^n - \{0\}$. Ciphers with low autocorrelation are known to be more secure.

Balance (regularity) is another important criterion which should be fulfilled by a Boolean function used in ciphering (see [17]). This means that each output bit (*0* or *1*) should appear an equal number of times for all possible values of inputs. The balance of a Boolean function is measured using its Hamming Weight (HW_f) and is defined as

$$HW_f = \frac{1}{2}(2^n - \sum_{x \in B^n} \hat{f}(x)) \cdot \tag{3}$$

Boolean function is balanced when its Hamming Weight is equal to 2^{n-1}.

Strict Avalanche Criterion (*SAC*) was first introduced by Webster and Tavares [14]. A Boolean function of n variables satisfies *SAC*, if complements of any of the n input bits result in changing the output bit with probability $1/2$. It means, that for each

of n-element vector c^n with only one the *i-th* bit of this vector equal to *1* (c_i^n) the following equation is satisfied

$$\sum_{x \in B^n} f(x) \oplus f(x \oplus c_i^n) = 2^{n-1} \cdot \tag{4}$$

The analysis of satisfaction of *SAC* for Boolean function f is measured by the distance $dSAC_f$, which is expressed by the equation

$$dSAC_f = \max_{1 \le i \le n} | 2^{n-1} - \sum_{x \in B^n} f(x) \oplus f(x \oplus c_i^n) | \cdot \tag{5}$$

The quality of block ciphers received with use of S-boxes is usually measured by criteria proper to Boolean functions. S-boxes are functions, which from n input bits generate k output bits. However, a Boolean function returns as output one bit. To use Boolean functions criteria to examine S-boxes, we need to transform all k bits output of an S-box into one output bit. After this modification, we obtain a new Boolean function which can be defined as f_β: $B^n \rightarrow B^1$, and expressed by the formula (see [3], [4], [9], [10])

$$f_\beta(x) = \beta_1 f_1(x) \oplus \beta_2 f_2(x) \oplus ... \oplus \beta_k f_k(x) \cdot \tag{6}$$

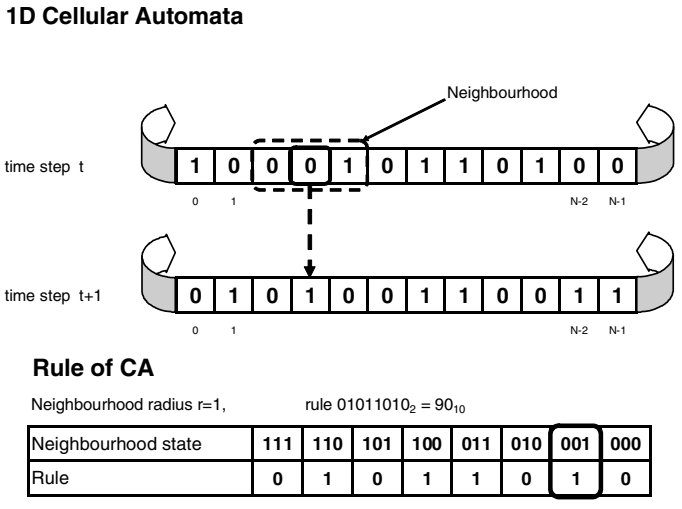
The new function $f_\beta(x)$ is a linear combination of k functions $f_i(x)$, where $i \le k$ and $\beta_i \in B^k$ are coefficients of the linear function. Each of functions $f_i(x)$ is a simple S-box (single-output S-box, a part of the n × k S-box mapped n input bits into the *i-th* of the k output bits). The relationship (vector (β_1, …, β_k)) between simple S-boxes is a result of the S-box table composition.

Cryptographical properties of S-boxes are calculated with use of the Boolean function $f_\beta(x)$. A linear combination of simple S-boxes in our computation was limited to trivial linear combination (where $\beta_i = 1$).

3 The Concept of Cellular Automata

One-dimensional (1D) CA is in the simplest case a collection of two-state elementary cells arranged in a lattice of the length N, and locally interacting in a discrete time t. For each cell i called a central cell, a neighborhood of a radius r is defined, consisting of $n_i = 2r + 1$ cells, including the cell i. When considering a finite size of CA, and a cyclic boundary condition is applied, it results in a circle grid (see Fig. 2).

It is assumed that a state q_i^{t+1} of a cell i at the time $t + 1$ depends only on states of its neighborhood at the time t, i.e. $q_i^{t+1} = f(q_i^t, q_{i1}^t, q_{i2}^t, ..., q_{in}^t)$, and a transition function f called a rule, which defines a rule of updating a cell i (see Fig. 2). A length L of a rule and a number of neighborhood states for a binary uniform CA is $L = 2^n$, where $n = n_i$ is a number of cells of a given neighborhood, and a number of such rules can be expressed as 2^L. Fig. 2 presents an example of the rule *01011010* (called also rule *90*) for $r = 1$. The length L of the rule consists of 8 bits.



Neighbourhood state	111	110	101	100	011	010	001	000
Rule	0	1	0	1	1	0	1	0

Fig. 2. 1D Cellular automata with neighborhood equal to 1

4 Construction of n × k CA-Based S-Boxes Generator

Cryptographic literature shows many examples and methods of searching S-box tables, not flexible structures requesting predefined sizes of memory (see [9], [3], [10], [14], [17]). In [9], [3], [10] authors treat the problem of designing S-box tables as a combinatorial optimization problem and apply different metaheuristics to search solutions in a huge space of S-box tables. Therefore, it is difficult to use them in new designed cryptographic algorithms, which request using dynamic S-boxes. The purpose of this study was to design flexible S-boxes, ready to use in cryptographic algorithms with dynamic S-boxes. It seems that CA is appropriate tool to design such S-boxes. Recently in [12], we reported a results of our study, where we analyzed the possibility to use CA as new kind of S-boxes (in the form not a determined tables, but some dynamically created structures), and we did it using the 8 × 8 S-boxes corresponding to AES S-boxes. We compared our results with S-box tables proposed by the Millan et al. [9], Clark et al. [3] and Nedjah & Mourelle [10]. Also in [13] we analyzed 6 × 4 CA-based S-boxes and compared them with corresponding to DES ones. Analyzed CA-based S-boxes are structures contained the initial CA configuration (100 cells), rule of CA, number of time steps (equal to 100) and the first n cells (in the initial configuration of CA) as inputs also the first k cells (in CA after predefined time steps) as outputs of CA-based S-box. So, we can conclude that CA-based S-box has a dynamical nature, but works on the predefined variables.

Now, we present a wide analysis of CA in the sense of dynamical the n × k CA-based S-boxes. Proposed dynamical CA-based S-boxes are fully randomly created structures. Each component of such a kind S-boxes is randomly selected from possible ranges (initial configuration, time steps, vector of inputs/outputs) or randomly selected from set of determined components (CA rule). So, the dynamic n × k

CA-based S-box can be considered as generator of CA-based S-boxes, where the seed of such a generator can be seen as CA consisted of the following elements:

- a range of a random number (N) of CA cells performing the role of background (an initial configuration of CA)
- a numbers of inputs (n) and outputs (k) used to construct a vector composed of a random numbers of CA cells performing the role of input/output of dynamic CA-based S-box (performing the role of arguments of S-box)
- a set of an appropriate CA rules from a rule of CA-based S-box will be randomly selected
- a range of a random number (T) of Time Steps during which CA will be evolved.

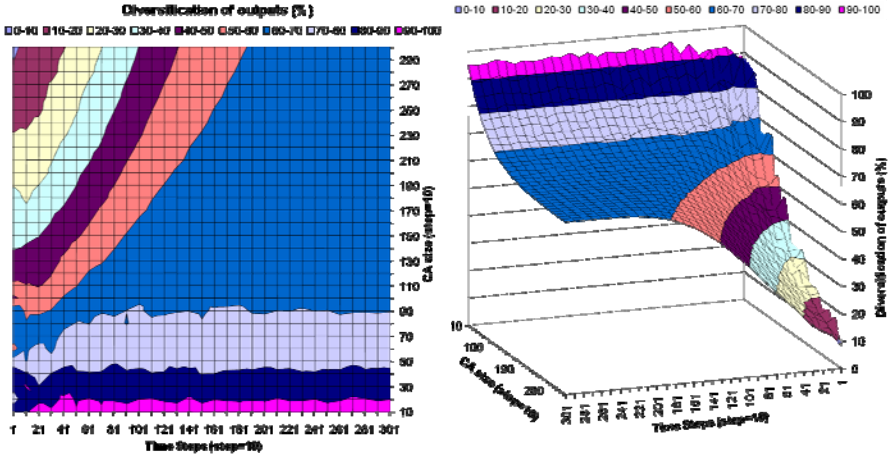
The first step in designing the dynamic n × k CA-based S-box is fixing the ranges from which parameters of S-box will be randomly selected. The first parameter (N) is a number of CA cells ($N \geq max|n, k|$). The second parameter is the vector composed of numbers of cells in which n inputs of S-box (in time step $t = 0$) and k outputs (in time step $t = T$) are arranged. Input/output cells are randomly arranged in CA cells. The third parameter is the set of CA rules. The fourth, the last parameter (T) is a number of time steps in which CA will be evolved.

5 Analysis of Parameters of CA-Based S-Boxes Generator

In this paper the dynamic 8 × 8 CA-based S-boxes will be analyzed. For this purpose we fixed the ranges for selecting CA size (N) as [10, 300] and for selecting time steps (T) of CA as [1, 300]. Not each CA rule is suitable to provide proper quality for CA-based S-box. From whole set of 256 elementary CA rules (CA, with neighborhood radius $r = 1$) we selected four rules {30, 86, 135 and 149} as the only proper ones for this purpose. Other single rules are too weak to be used in CA-based S-boxes, because they are characterized by not high enough non-linearity and balance, also not low enough autocorrelation and $dSAC_f$. Selected the best rules change CA cells in time step t into cells in time step $t + 1$, as follows:

Rule 30: $q_i^{t+1} = q_{i-1}^{\ t} \oplus (q_i^{\ t} \vee q_{i+1}^{\ t})$,
Rule 86: $q_i^{t+1} = (q_{i-1}^{\ t} \oplus q_i^{\ t}) \vee q_{i+1}^{\ t}$,
Rule 135: $q_i^{t+1} = \neg[q_{i-1}^{\ t} \oplus (q_i^{\ t} \vee q_{i+1}^{\ t})]$,
Rule 149: $q_i^{t+1} = \neg[(q_{i-1}^{\ t} \oplus q_i^{\ t}) \vee q_{i+1}^{\ t}]$.

Selected rules are in the 3-rd class of CA in Wolfram's classification (see [16]). It means that they have property of randomness. CA managed by one of these rules for a random initial configuration generates during the time steps random bit sequences for each of CA cell. For selected rules values of computed cryptographic criteria like the non-linearity, autocorrelation, balance and distance to fulfillment the SAC are better than for another rules (see also [12]).



(a) **(b)**

Fig. 3. Diversification of outputs (in %) of dynamical 8 × 8 CA-based S-boxes for lattices composed of CA cells with a number of cells from the range [10, 300] and time steps evolution from the range [1, 300], 2D view (a) and 3D view (b)

The first experiment of our study concerns the analysis of bijectivity [8] of dynamic 8 × 8 CA-based S-box. We analyzed diversification of 8 outputs arranged by randomly chosen CA cells (vector inputs/outputs of S-box). For the lattices of the size ranging in [10, 300] CA cells and [1, 300] time steps of evolution CA, we calculated the average (from 10000 randomly selected initial configurations of CA) diversification of dynamic CA-based S-box outputs. It means that 930 points (single CAs) were calculated in each lattice. For each of 10000 CA components of vector (of 8 input/output cells) were randomly selected and arranged in random CA cells. Results of these calculations are presented in Fig. 3. Fig. 3a shows percentage diversification of outputs in dynamical CA-based S-box as 2D form of a diagram. The same property is also presented in Fig. 3b as a surface (3D form) shaped by such properties in a number of CA cells and time steps of CA.

Results of those experiments show that exists a relationship between parameters of CA size and time steps. Preliminary obtained average diversifications of S-box outputs are quite high (more than 60%) and stable, if *value of time steps is not lower than CA size* (see Fig. 3). This rough approximated relationship between the number of CA cells (N) and the number of time steps of CA evolution (T) is good enough for preliminary analysis of our approach to dynamical S-boxes. This relationship is expressed by the formula

$$T \geq N. \tag{7}$$

Results of analysis cryptographic criteria for CA-based S-boxes generated under above relationship between number of CA size (N) and time steps (T) shows that the best and stable values of non-linearity is not lower than 80, autocorrelation is not higher than 120, Hamming Weight is not lower than 80 (and in most is higher than 100) and distance to fulfillment SAC is not higher than 40.

6 Analysis of Cryptographical Properties of Generated CA-Based S-Boxes

The analysis presented in the previous section shows that dynamic CA-based S-boxes give a stable and the best results for time steps (T) not lower than the number of CA cells (N). Under this condition we constructed dynamical 8 × 8 CA-based S-boxes, where CA size ranging in [10, 300], time steps ranging in [1, 300], and (T ≥ N). CA rules are selected from the set {30, 86, 135, 149}. For each initial configuration of CA, a vector consisted of the 8 input/output cells of CA was randomly selected. Such a composed generator of S-boxes can generate huge number of different S-boxes, this number equals to $\sum_{n=10}^{N}\sum_{t\geq n}^{T} 4*2^n*t*C_n^8*8!$. Such a huge space of possible to generate CA-based S-boxes makes ciphering with use of this generator much safer.

Also, we analyzed cryptographical properties (as non-linearity, autocorrelation, balance expressed by Hamming Weight and $dSAC_f$) of such S-boxes. Results of this analysis are presented in diagrams (Fig. 4).

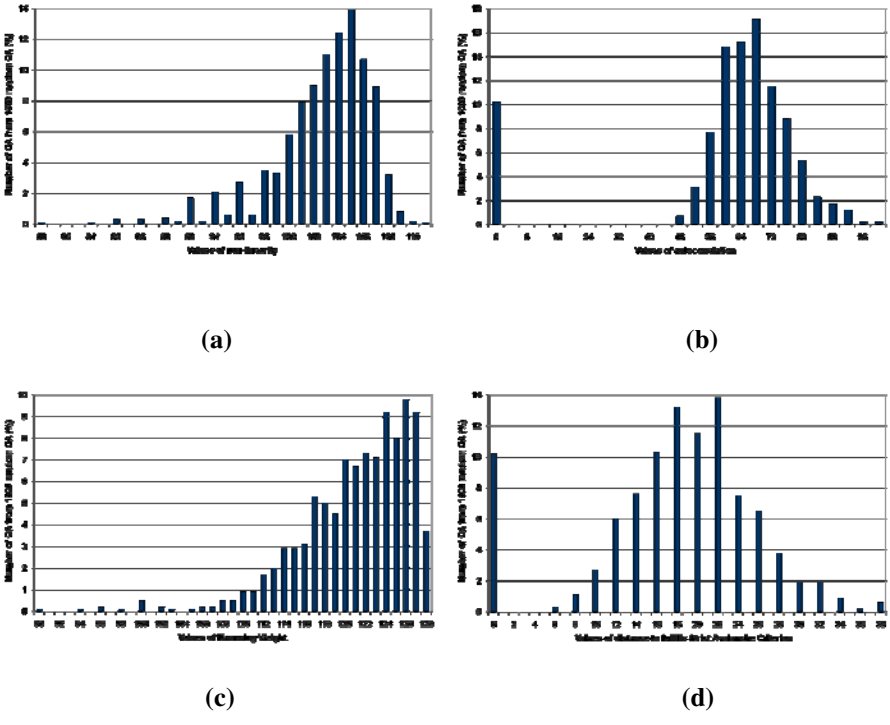


Fig. 4. Percentage of CA corresponding to non-linearity (a), autocorrelation (b), Hamming Weight (c) and distance to fulfillment of the Strict Avalanche Criterion (d) of dynamical 8 × 8 CA-based S-boxes for 1000 random CA, with CA size from the range [10, 300] and time steps from the range [1, 300]

For the 8 × 8 S-boxes theoretical values of N_f ranged in [0, 128], where the best (ideal) value is equal to 128, AC_f ranged in [0, 256], where the best (ideal) value is equal to 0, HW_f ranged in [0, 128], where the best (ideal) value is equal to 128, and $dSAC_f$ ranged in [0, 128], where the best (ideal) value is equal to 0.

One can see that values of cryptographic criteria corresponding to non-linearity (Fig. 4a), autocorrelation (Fig. 4b), Hamming Weight (Fig. 4c) and $dSAC_f$ (Fig. 4d) are not so far from the best theoretical values related to such kind of 8 × 8 S-boxes. Moreover, in the set of selected CA-based S-boxes exist more than 10% with the best of possible value (equal to 0) for autocorrelation (see Fig. 4b) and $dSAC_f$ (see Fig. 4d). Also, there exists 4% perfectly balanced (with value 128) CA-based S-boxes (see Fig. 4c).

Non-linearity of this randomly selected CA-based S-boxes is high, and for ~ 85% solutions is not lower than 100. The most frequently observed value is 105 (for 14% solutions) and the best obtained value of non-linearity is 111, obtained for few CA-based S-boxes.

Summarizing, obtained results are very promising, values of cryptographic properties are good and quality of such composed CA-based S-boxes is high. So, application of dynamical CA-based S-boxes makes the cryptosystems with its use more efficient and stronger.

7 Conclusions

The paper presents an idea of creating S-boxes using CA-based approach. Classical S-boxes based on tables are fixed structure constructions. We are interested in creating CA-based S-boxes, which are dynamical structures. CA from input block of bits generates output block of bits and is evaluated by the same examination criteria as the traditional S-box.

Conducted experiments have shown that the dynamical n × k CA-based S-boxes are characterized mostly, by a high non-linearity and balance expressed by Hamming Weight, and also low autocorrelation and distance to fulfillment of the Strict Avalanche Criterion.

Dynamic CA-based S-box is high quality generator of CA-based S-boxes, and the space of such generated S-boxes is high, in opposite to stable classical S-box tables, which in general are characterized by lower quality than CA-based S-boxes.

Also, dynamic version of CA-based S-boxes is easy to use in cryptographical systems and supply high space of possible S-boxes; moreover, they provide high quality in the sense of cryptographic properties, what make safe ciphering with its use.

Acknowledgments. This work was supported by Polish Ministry of Science and Higher Education as the grant No. N N519 388036.

References

1. Adams, C., Tavares, S.: Gooad S-boxes are easy to find, Advances in cryptology. In: Brassard, G. (ed.) CRYPTO 1989. LNCS, vol. 435, pp. 612–615. Springer, Heidelberg (1990)

2. Albert, J., Culik II, K.: A simple universal cellular automaton and its one-way and totalizing version. Complex Systems 1, 1–16 (1987)
3. Clark, J.A., Jacob, J.L., Stepney, S., The Design, S.: The Design of S-Boxes by Simulated Annealing. New Generation Computing 23(3), 219–231 (2005)
4. Dowson, E., Millan, W., Simpson, L.: Designing Boolean Functions for Cryptographic Applications. Contributions to General Algebra 12, 1–22 (2000)
5. Federal Information Processing Standards Publication, FIPS PUB 46-3, DES (1999), http://csrc.nist.gov/publications/fips/fips46-3/fips46-3.pdf
6. Federal Information Processing Standards Publications, FIPS PUBS 197, AES (2001), http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf
7. Feistel, H.: Cryptography and Computer Privacy. Scientific American 228(5), 15–23 (1973)
8. Millan, W.: How to Improve the Non-linearity of Bijective S-boxes, LNCS, vol. In: Maekawa, M., Belady, L.A. (eds.) IBM 1980. LNCS, vol. 143, pp. 181–192. Springer, Heidelberg (1982)
9. Millan, W., Burnett, L., Carter, G., Clark, A., Dawson, E.: Evolutionary Heuristics for Finding Cryptographically Strong S-Boxes. In: Varadharajan, V., Mu, Y. (eds.) ICICS 1999. LNCS, vol. 1726, pp. 263–274. Springer, Heidelberg (1999)
10. Nedjah, N., de Macedo Mourelle, L.: Designing Substitution Boxes for Secure Ciphers. International Journal Innovative Computing and Application 1(1), 86–91 (2007)
11. Smith III, A.R.: Simple computation-universal cellular spaces. Journal ACM 18, 339–353 (1971)
12. Szaban, M., Seredynski, F.: Cryptographically Strong S-Boxes Based on Cellular Automata. In: Umeo, H., Morishita, S., Nishinari, K., Komatsuzaki, T., Bandini, S. (eds.) ACRI 2008. LNCS, vol. 5191, pp. 478–485. Springer, Heidelberg (2008)
13. Szaban, M., Seredynski, F.: Cellular Automata-based S-Boxes vs. DES S-Boxes. In: Malyshkin, V. (ed.) PaCT 2009. LNCS, vol. 5698, pp. 269–283. Springer, Heidelberg (2009)
14. Webster, A.F., Tavares, S.E.: On the Design of S-Boxes. In: Williams, H.C. (ed.) CRYPTO 1985. LNCS, vol. 218, pp. 523–534. Springer, Heidelberg (1986)
15. Wolfram, S.: Universality and complexity in cellular automata. Phisica D 10, 1–35 (1984)
16. Wolfram, S.: A New Kind of Science. Wolfram Media Inc., Champaign (2002)
17. Youssef, A., Tavares, S.: Resistance of Balanced S-boxes to Linear and Differential Cryptanalysis. Information Processing Letters 56, 249–252 (1995)

Detection of LiveLock in BPMN Using Process Expression

Nasi Tantitharanukul and Watcharee Jumpamule

Faculty of Science, Department of Computer Science,
ChiangMai University, ChiangMai, Thailand 50200
g510531113@cm.edu, sccsi003@chiangmai.ac.th

Abstract. Although the Business Process Modeling Notation (BPMN) is a popular tool for modeling business process in conceptual level, the result diagram may contain structural problem. One of the structural problems is livelock. In this problem, one token proceeds to end event, while other token is still in process with no progression. In this paper, we introduce an expression liked method to detect livelock in the BPMN diagram. Our approach utilizes the power of the declarative ability of expression to determine all of the possible process chains, and indicate whether there are livelock or not. As a result, we have shown that our method can detect livelock, if any.

Keywords: BPMN, Livelock, Automata, Expression.

1 Introduction

Although the information technology (IT) becomes the factor of success for the firm[1], the development still faces problems. According to survey conducted by the Standish Group[2], only 29% of software projects succeeded. The primary reason of these failures is poor conceptual modeling. One tool which is effective for modeling business process at the conceptual level is the Business Process Modeling (BPM). It is a tool used to characterize the identification and the specification of the business processes. However, the general modeling includes the arbitrariness and lacks the strictness[3]. For example, the congestion of business flows. If the congestion is not properly controlled, the software may not meet the requirements. Moreover, the congestion can lead to the structural problem such as livelock[4]. A livelock problem is a infinite excution of process. In this problem, some of the process may run successfully but some of the processes trap in a endless loop of excution. For example, a web browser loads a page called A that redirects to a page B that redirects back to A[5]. Thus, before the model being implemented, we need to verify formally them.

There are many approaches being applied to verify the BPM. In [6], a Petri net based method to verify and to validate BPMN has been proposed. This method can detect various types of error which includes livelock. van Dongen and others have proposed a method to detect livelock by using concept of footprints in [7]. The footprint analysis is a technique to analysis the compatibility of business

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 164–174, 2010.

process diagram by taking the set of previous process and the set of afterward process into account.

In our previous work[8], we have proposed a method to detect deadlock and multiple termination by transforming the BPMN diagram into process automata and verify them. In this paper, we propose a method to detect livelock in BPMN. The working strategy of this work is to convert the BPMN into a process expression and try to find a certain property which is related to livelock. The process expression is an other way to represent the BPMN in algebraic approach. Since the actual object which is really important to the business process management is the business activity and the object to control the flow, we only takes functional feature of BPMN model, e.g. gate and process, into account but not the non-functional feature, e.g. association, artifacts, and organizational feature, e.g. pool and lanes. In this work, we assume that there is no other structural problem such as deadlock and multiple terminations.

In the next section, we give an overview over BPMN. Section 3 introduces preliminaries. In this section, the livelock and the process expression are also formally defined. A method to detect livelock in BPMN model using process expression is presented in Section 4. The experiment and the result are presented Section 5. In the last section, we conclude this paper and discuss the future work.

2 BPMN

The Business Process Modeling Notation (BPMN) is a set of graphical objects and rules that is standardized by OMG. The aim of the model is to fill the gap between the business design and its implementation. Due to its ease to understand, the BPMN has become de facto for capturing business processes[9].

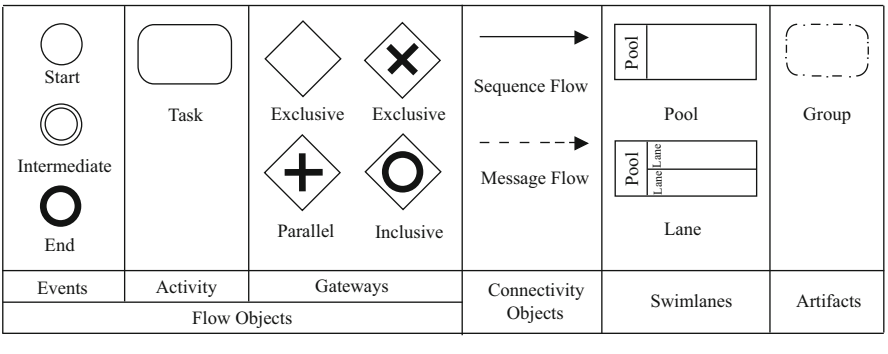


Fig. 1. BPMN's components

As described in the specification of BPMN[10] and shown in Fig.1, there are four basics categories of elements of BPMN: flow objects, connecting objects, swimlanes, and artifacts.

- Flow objects - objects used for representation the behavior of a business process. There are three types of flow objects:
  - Events - used to express something that happens in business process,
  - Activity - used for representation the work that the firm performs,
  - Gateway - used to control the divergence and convergence of sequence flow. There are three types of gateways:
    - Exclusive gateway - used to restrict the flow to chose only a single path,
    - Parallel gateway - used to restrict the flow to chose at least one path, and
    - Inclusive gateway - used to restrict the flow to chose all of the path(s).
- Connecting objects - objects used to connect the flow objects to each other,
- Swimlanes - objects used to group the modeling elements, and
- Artifacts - objects used to provide additional information.

3 Preliminaries

In this section, we introduce basic facts and notation used in this research.

As mentioned in Section 2, the BPMN is considered as a group of set of its components and a relation among them. We follow our previous work[8]. Formally, the BPMN can be defined as follows

Definition 1 (Business Process Modeling Notation). *A BPMN model is a tuple* $B = (F, E, E^s, E^e, A, G, G^p, G^e, T)$ *where*

- F *is a finite set of flow objects in BPMN process which can be partitioned into events* E*, activities* A *and gateways* Γ*,*
- E *is a finite set of events which can be partitioned into start event* E^s*, and end event* E^e*,*
- A *is a finite set of activities in BPMN process,*
- G *is a finite set of gateways in BPMN which can be partitioned into inclusive gateway* G^i*, parallel gateway* G^p *and exclusive gateway* G^e*, and*
- *Transition* $T \subseteq F \times F$ *is a finite set of sequence flows connecting objects.*

The processing of BPMN model utilizes the concept of "Token" that will traverse the sequence flow. A sequence of process which is occurred during the processing is called a process chain. Since the *gateway* control the *flow* and the *event* mark the happening of the incident, the actual process that will be used during the processing is *activity*.

Definition 2 (Process Chain). *Process Chain is a sequence of symbols from* A *in a BPMN model. Formally*

$$a_1 a_2 \dots a_k| \text{ a constant } k \geq 0 \text{ and } a_i \in A, i = 1, 2, \dots, k$$

Since we have already defined BPMN and process chain, we now investigate further into their properties. We define three operations on process chain and use them to study properties of process chain.

Definition 3. *Let C_1 and C_2 be set of process chains. We define the operations as follows*

- *Merge* : $C_1 \oplus C_2 = \{c | c \in C_1 \text{ or } c \in C_2\}$,
- *Concatenation* : $C_1 \cdot C_2 = \{c_1 c_2 | c_1 \in C_1 \text{ and } c_2 \in C_2\}$, *and*
- *Repeated* : C_1^* *is either*
  1. $c \in C_1$,
  2. ϵ, *or*
  3. $C_1 \cdot C_1^*$.

The Merge operation is similar to the union operation except that the Merge operation represents the gateway of BPMN. In this research, we do not take the semantic distinction of gateway into account. Therefore, we only mark at where the consecutive sequence can be more than one. The Concatenation operation attaches the process chain from C_1 to the front of the process chain from C_2.

The Repeated operation is the most important operation in this research. It concatenates zero or more process chain together from the same set of process chain. The ϵ represents the empty process, a sequence of length zero. This operation is later used in the detection of livelock.

The closure properties of the process chain under these operations can be proofed by construction approach. The Merge operation can be constructed by using the gateway. The Concatenation operation is the flow objects connecting by a sequence flow. The Repeated operation is the connection between the antecedent objects and the following objects.

Now, we switch our representation from a graphical approach to a algebraic approach: the process expression. This notation involves a combination of symbols and operations to represent the set of process chain. In addition, the process expression can define the set of process chains that really happens clearer than the diagram represented by BPMN.

Definition 4 (Process Expression). *Say that P is a process expression if P is either*

1. $a \in A$,
2. ϵ,
3. $(P_1 \oplus P_2)$, *where P_1 and P_2 are process expressions,*
4. $(P_1 \cdot P_2)$, *where P_1 and P_2 are process expressions, or*
5. (P_1^*), *where P_1 is a process expression,*

Like other algebras, the process expression also has an order associated with operation called the precedence. For the process expression, the precedence of the process expression is as follows: (), $*$, $\cdot$, and $\oplus$.

We may now formally define the livelock. A livelock is a situation where there exist a cycle(s) in BPMN transition as shown in Fig.5. The problem lies in the intersection between the cycle and the outside; i.e., g_2. For instance, as shown in Fig.5, if gate g_4 chooses both a_5 and a_6, the a_6 will proceed to the e_2 but a_5 will remain in the processing. In worst case scenario, only a_5 is chosen and is fired continuously without any progress. In other words, the cycle in the BPMN diagram can consume the infinite number of process.

Definition 5 (Livelock). *A BPMN model is considered to have livelock iff it can consume infinite number of process chain.*

4 Livelock Detection Method

After defining preliminaries, we show that the livelock can be matched with property in process expression. Then, we show that any BPMN diagram can be converted to process expression. We introduce the conversion algorithm from BPMN to process expression. Using this algorithm, we can detect the livelock in any BPMN diagrams.

We first show that the livelock is equivalent to Repeated operation.

Theorem 1. *If a process expression contains repeated operation, then it also contain livelock.*

Proof. Let $P^{'}$ be a process expression where $P^{'}$ contains repeated operation. We call this process expression P^{*}. By Def.3, we have

$$J \in P^{*} | J = \underbrace{P \cdot P \cdot P \cdot \ldots \cdot P \cdot P \cdot P}_{k} \text{ where } k \geq 0 \tag{1}$$

Let c be an arbitrary process chain which can be generated from P. So,

$$c = a_1 a_2 a_3 \ldots a_{j-2} a_{j-1} a_j | \text{ a constant } j \geq 0$$

From Equa.1, J generates a process chain $c^{'}$ where

$$c^{'} = \underbrace{c \cdot c \cdot c \ldots c \cdot c \cdot c}_{k}$$

From $k \geq 0$, if we consider $k = \infty$, for each process a_j in c will be executed infinitely in $c^{'}$. By the Def.5, $c^{'}$ causes livelock. Thus, P^{*} causes livelock. □

Next, we show the conversion of BPMN diagram to process expressions.

Theorem 2. *If a set of process chain is described by a BPMN diagram, then it also can be described by process expression.*

Proof. What we need to do is to find a process expression capable of describing all the possible process chain from start event to end event. In order to fulfill our goals, we have to describe a procedure to convert any BPMN diagram into process expression without losing any semantic meaning.

We break this into two steps. First, we transform the BPMN diagram into a temporary form called Generalized Business Process Modeling Notation. With this form, we have the process expression on the node instead of BPMN's flow object. Then, we recursively reduce the number of node until there are only three nodes left in the diagram: the start event, the end event, and the process expression represented the entire diagram. The result process expression from

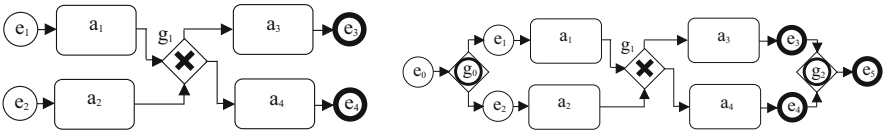
the second step is the concatenation of several process expressions, and hence the process itself is process expression by the Def.4.

A Generalized Business Process Modeling Notation (GBPMN) is a diagram whose nodes are labeled with the process expression. We formally define the GBPMN as follow.

Definition 6. *A GBPMN model is a tuple* $G = (R, E, e^s, e^e, \Gamma, H)$ *where*

- R *is a finite set of process expression,*
- E *is a finite set of events,*
- $e^s \in E$ *is start event,*
- $e^e \in E$ *is end event,*
- Γ *is a finite set of gateways, and*
- *Transition* $H \subseteq R \cup \Gamma \times R \cup \Gamma$ *is a finite set of sequence flows connecting objects.*

We can easily convert any BPMN into GBPMN form. At the beginning, by Def.4, the set of process expression, R, is equal to the set of activity in the BPMN. Then, if the existing diagram has only one single start event and one single end event, it is already in GBPMN form. If not, we have to convert them by using the following method. We first add a new start event and a new end event before connecting these events to the existing diagram by using inclusive gateway which is capable of capturing whether they simultaneously start or not. The example is shown in Fig.2. Moreover, Γ is equal to the set of gateways in the BPMN diagram.



(a) The BPMN before transformation.

(b) The result GBPMN.

Fig. 2. An example of BPMN diagram and the derived GBPMN diagram

The reduction of object is the most important step in conversion of the BPMN into process expression. The concept of the reduction method is to choose any transition, and replaces with a new process expression which can describe its participants without losing semantic definition based on the participants. We first select the transition in GBPMN called t_{rip}. For all the transition in GBPMN, there always have two participants: the antecedent called O_i, and the descendant called O_j. To be more precise, both of the participants must be process expressions. Based on these participants, there are three possible scenarios to consider:

- Both O_i and O_j are activities. We first remove the ripped transition, and insert a new process expression concatenating O_i and O_j. Suppose that t_{rip} is (O_i, O_j). In this case, the new process expression is labeled $O_i \cdot O_j$.

– One of O_i or O_j is gateway and the other is activities. In this case, we have to consider the structure of the transition. Since we do not take the semantic definition of gateway into account, what we can do is to mark the possible path of the same direction with merge operation. Moreover, the paths have to be separated from the consecutive object. When O_i is gateway, we have to put the open bracket before we merge the paths. And, when O_j is gateway, we have to merge the path before we put the closed bracket. Then, we concatenate the result process expression with the consecutive object. There are two type of structure: the forward direction and the backward direction.
  - For the forward direction, the direction of the ripped transition goes direct to the end event. In this direction, we ignore the gateway, and merge the paths to construct the process expression. Suppose that O_i is gateway and $(O_i, O_a), \ldots, (O_i, O_k) \in H$. In this case, we have a new process expression labeled$(O_a \oplus \ldots \oplus O_k)$.
  - For the backward direction, the direction of the ripped transition is in the opposite with the forward direction, and create the loop in the diagram. The O_i of the backward transition always is O_j of the forward direction transition. So, after a backward path ends, another forward path begins. Moreover, the loop can be repeated as much as possible. Therefore, we begin the process expression by the merge of the backward path, we concatenate them with the merge of forward path, and we apply the repeated operation. Suppose that O_i is gateway, $(O_i, O_a), \ldots, (O_i, O_k) \in H$, and $(O_l, O_i), \ldots, (O_z, O_i) \in H$. In this case, we have a new process expression labeled $((O_a \oplus \ldots \oplus O_k) \cdot (O_l \oplus \ldots \oplus O_z))^*$.
– Both O_i and O_j are gateways. In this case, we follow the second scenario except that we use ϵ as the process expression for this transition.

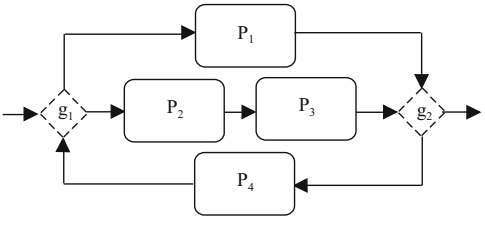


Fig. 3. GBPMN's components

When we combine the process expressions, we have the following formula and its BPMN diagram shown in Fig.3. According to the scenarios, there are three possible expressions.

$$\left(\underbrace{P_1}_{1} \oplus \overbrace{P_2 \cdot P_3}^{2} \right) \cdot \underbrace{(P_4 \cdot (P_1 \oplus P_2 \cdot P_3))^*}_{3}$$

The first part represents the combination of the forward direction transition. The second part represents the concatenation of two activities. Finally, the third part represents the combination of the backward direction transition. To prevent confusion, we only apply each part of our formula in expression only where the ripped transition matches the scenario. For example, we only concatenate $(P_4 \cdot (P_1 \oplus P_2 \cdot P_3))^*$ when t_{rip} is a backward transition.

Now, we show how to convert a GBPMN G into a process expression.

Algorithm 1. *CONVERT(G)*

1. *If the number of object in G are three, then G must consist of a start event, an end event, and a process expression whose labeled describe the all possible process chain from the start event to the end event, then return the process expression.*
2. *If the number of object in G are more than three, we have to select an arbitrary t_{rip} from G where $t_{rip} = (O_i, O_j)$ such that either O_i or O_j is not start event or end event, and we replace O_i and O_j with a new process expression labeled with R where*

$$P = (P_1 \oplus P_2 \cdot P_3) \cdot (P_4 \cdot (P_1 \oplus P_2 \cdot P_3))^*$$

   Let G' be the GBPMN $(R, E, e^s, e^e, \Gamma, H)$ that has removed O_i and O_j, and replaced with P.
3. *return CONVERT(G').* □

Next, we show that the algorithm work correctly.

Claim 1. *For any GBPMN G, CONVERT(G) is equivalent to G.*

We proof this claim by induction on the number of object in diagram, k.

Proof. **Basis:** Prove this claim true for $k = 3$. If G has only three objects, since the diagram must contain e^s and e^e, it can have one single process expression called P_1. Seeing that P_1 describes all the possible path from start event to end event and by line 1 in CONVERT, the P_1 is equivalent to G.

Induction step: Assume that the claim is true for $k - 1$ objects and use this assumption to prove that the claim is true for k objects.

We insert a new object in between P_i and P_j. Three possibilities occur during the insertion of a new object. Either the object is inserted next to the activities, the object is inserted in the forward transition, or the object is inserted in the backward transition. For each direction, CONVERT(G) describes all the possible process chain from P_i to P_j as mentioned in Theorem 2. Thus, CONVERT(G) is equivalent to G. □

From Theorem 1 and Theorem 2, we can detect the livelock in BPMN diagram by converting the BPMN into process expression and determine whether it contain repeated operation or not.

Corollary 1. *If a BPMN diagram G contains livelock, then the process expression $G^{'}$ which is converted from G contains repeated operation.*

5 Experiment and Result

In this section, we discuss the use of our livelock detection method with some examples. In this work, we take requirements of the online shopping system into account. The system of interest begins when the customer search the merchandises. Then, the customer may either add the merchandises to the shopping cart or remove the merchandises from the shopping cart. After the customer finish the shopping process, the customer has to pay for the goods by credit card.

5.1 Correct BPMN

In this case, we apply our method on the BPMN diagram of the online shopping system shown in Fig.4. In this diagram, it is clearly that there is no backward direction transition. Therefore, the possibility that this diagram contain livelocks will never happen. We can derive the process expression as $(a_1 \cdot a_2) \cdot (a_3 \oplus a_4) \cdot (a_5 \cdot a_6)$. In this example, there is no repeated operation in the process expression. Thus, by Corollary.1, this diagram contains no livelock.

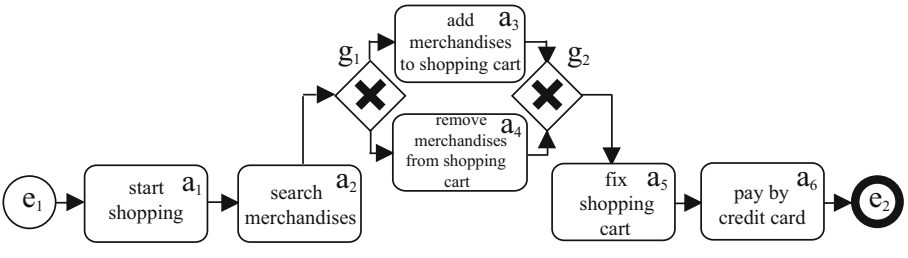


Fig. 4. The BPMN diagram with no livelock

5.2 BPMN with Livelock

In this case, we apply our method on the BPMN diagram of the online shopping system shown in Fig.5. In this diagram, the system allows the customer either to fix shopping cart or to continue shopping which may result in the miscalculation of product's quantity and total price. As shown in Fig.5, the process a_5 go back to gate g_1 which may cause livelock. Therefore, this diagram contains a potential livelock. We can derive the process expression as $a_1 \cdot ((a_2 \cdot (a_3 \oplus a_4)) \cdot (a_5 \cdot (a_2 \cdot (a_3 \oplus a_4)))^*) \cdot a_6 \cdot a_7$. In this example, there is repeated operation in the process expression. Thus, by Corollary.1, this diagram contain livelock.

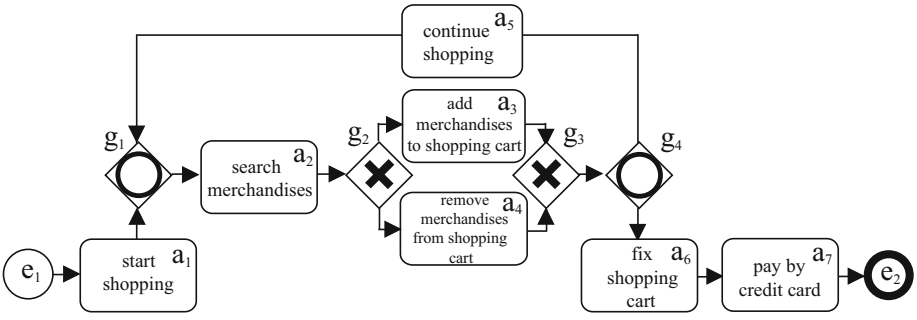


Fig. 5. The BPMN diagram with livelock problem

6 Conclusion and Future Work

In this paper, we have presented an approach to detect livelock in BPMN diagram. Our approach makes use of the declarative power of process expression to identify the livelock in the diagram. The proposed method begin with the transformation of BPMN to GBPMN in order to generalize the structure of the diagram. Then, we convert the diagram in GBPMN form to the process expression form. When the process expression contains Repeated operation, the BPMN diagram have livelock. The result shows that our method can detect the livelock in the diagram.

For the future work, we intend to add semantic definition of gateway into process expression since some combination of gateway can cause other type of structural problem such as deadlock.

References

1. Porter, E., Millar, V.: How information gives you competitive advantage. Harvard Bus. Rev. 63, 149–162 (1985)
2. Standish Group International: Third Quarter Research Report. Technical report, The Standish Group International, Inc., Boston (2004)
3. Morimoto, S.: A Survey of Formal Verification for Business Process Modeling. In: Bubak, M., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2008, Part II. LNCS, vol. 5102, pp. 514–522. Springer, Heidelberg (2008)
4. Ouyang, C., Dumas, M., van der Aalst, W., ter Hofstede, A.: From Business Process Models to Process-oriented Software Systems: The BPMN to BPEL Way (2006)
5. Ho, A., Smith, S., Hand, S.: On deadlock, livelock, and forward progress. Technical Report, University of Cambridge, Computer Laboratory (May 2005)
6. Raedts, I., Petkovic, M., Usenko, Y., Werf, J.M., Groote, J.F., Somers, L.: Transformation of BPMN models for Behaviour Analysis. In: MSVVEIS, Madeira, Portugal, pp. 126–137 (2007)
7. van Dongen, B.F., Mendling, J., van der Aalst, W.M.P.: Structural Patterns for Soundness of Business Process Models. In: EDOC 2006: Proceedings of the 10th IEEE International Enterprise Distributed Object Computing Conference, pp. 116–128 (2006)

8. Tantitharanukul, N., Sugunsil, P., Jumpamule, W.: Detecting Deadlock and Multiple Termination in BPMN Using Process Automata. In: ECTICON 2010: Proceedings of the 7th ECTI International Conference on Electrical Engineering/Electronics, Computer, Telecomunications and Information Technology, Chiang Mai, Thailand (2010)
9. Dijkman, R., Dumas, M., Ouyang, C.: Formal semantics and automated analysis of BPMN process models. Technical report, Queensland University of Technology, Queensland (2007)
10. Object Management Group (OMG): Business Process Modeling Notation (BPMN) Specification, Final Adopted Specification. Technical report, Object Management Group, Needham (2006)

The Effect of Background Traffic Packet Size to VoIP Speech Quality

Tuul Triyason, Prasert Kanthamanon, Kittipong Warasup, Siam Yamsaengsung, and Montri Supattatham

School of Information Technology, King Mongkut's University of Technology Thonburi, Pracha-utid Road, Bangmod, Toongkru, Bangkok, Thailand
51501701@st.sit.kmutt.ac.th,
{prasert,kittipong,siam,montri}@sit.kmutt.ac.th

Abstract. VoIP is gaining acceptance into the corporate world especially, in small and medium sized business that want to save cost for gaining advantage over their competitors. The good voice quality is one of challenging task in deployment plan because VoIP voice quality was affected by packet loss and jitter delay. In this paper, we study the effect of background traffic packet size to voice quality. The background traffic was generated by Bricks software and the speech quality was assessed by MOS. The obtained result shows an interesting relationship between the voice quality and the number of TCP packets and their size. With the same amount of data smaller packets affect the voice's quality more than the larger packet.

Keywords: VoIP, voice quality, MOS, packet size.

1 Introduction

Voice over Internet Protocol (VoIP) is one of the technologies that allow you to make a phone call over the IP network. VoIP is gaining popularity among the small and medium business because it has an advantage in cost saving when compare to the regular landline and mobile service. While, good voice quality is one of the challenging tasks in deployment plan. VoIP is a delay sensitive application and sensitive to many degradation factors such as packet loss and delay jitter because most deployment of VoIP is built over the data network. This make VoIP traffic always coexisting with other types of traffic, such as HTTP or FTP which have various packet sizes.

In this work, the Mean Opinion Score (MOS) [1] and R-Factor [2] are used to measure the VoIP's voice quality. The research aims to analyze the relation between the packet sizes of background traffic and MOS. In particular, we analyze how the quality of VoIP is affected when there is TCP background traffic with different packet sizes loading in the network. We show that the small packet size of TCP background traffic produces more packet loss than normal packet size and It also drops MOS scores.

This paper is organized as follows. In Section 2, we describe the speech quality assessment methodology. In Section 3, we describe about the analysis of background traffic and speech quality degradation. In Section 4, we present the experimental setup

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 175–182, 2010.

and results. We give further discussion in Section 5. In the last section, we draw the conclusions and the direction of future work.

2 The Speech Quality Assessment

The Mean Opinion Score (MOS), defined by ITU-T P.800, is accepted as a voice quality assessment. MOS is generated by a number of listeners against the hearing speech quality. A listener is given a speech quality rating by using the scale from 1 to 5. However, the standard test is taken a long time and not reliable to test.

The Emodel was originally developed by ETSI ad hoc group as a voice transmission quality prediction. The model is described in detail in ITU Recommendations G.107 [3] and G.108 [4]. The model is used to estimate the quality of conversation from "mouth to ear" and represents in the R-value. The R-value has a typical range of 0-100. However, the R-value that is below 50 is unacceptable.

Table 1. Provisional guide for the relation between R-value and user satisfaction from ITU-T

***R*-value**	**MOS**	**User satisfaction**
90	4.34	Very satisfied
80	4.03	Satisfied
70	3.60	Some users dissatisfied
60	3.10	Many users dissatisfied
50	2.58	Nearly all users dissatisfied

The Emodel can be translated into MOS. The provisional guide of the relation between R-value and user satisfactions from ITU-T are shown in Table 1. The Emodel merged the effects of several transmission parameters that impact the conversational quality into R-value.

$$R = R_0 - Is - Id - Ie + A \tag{1}$$

The equation (1) shows the R-value calculation. Where, R_0 represents the basic signal-to-noise ratio. The factor Is is the results of all impairments of the voice signal. The factor Id and Ie are the impairments caused by delay and low-rate codec. The factor A represents the compensation of impairment factors when It is convenient for the user to make the phone call. ITU-T G107 document was shown the equation for convert between R-value to MOS as follow

$$\begin{aligned} &\text{For R} < 0\text{: MOS} = 1 \\ &\text{For } 0 < \text{R} < 100\text{: MOS} = 1 + 0.035R + R(R-60)(100-R)7\cdot 10^{-6} \\ &\text{For R} < 100\text{: MOS} = 4.5 \end{aligned} \tag{2}$$

3 Analysis of Background Traffic and Speech Quality Degradation

3.1 Speech Quality Degradation Factor

From Section 2, it is clear that the impairment factor that has the greatest effect on voice quality is packet loss and delay. The impairment factors *Id* and *Ie* represent the degradation on speech quality due to delay and packet loss in VoIP. Most of packet losses come from two components. The first component comes from transmission loss and the second component comes from the drops packet by buffer overflow from congestion control scheme. In low packet loss rates, a burst distribution gave a higher subjective quality than a non bursty distribution [5]. The impact of packet loss on speech quality depends on several factors e.g. loss pattern, codec type, packet loss size and location of loss [6][7][8].

3.2 Impact of Background Traffic

It is clear that packet delay and packet loss have an important role on VoIP speech quality. Most of packet delay and losses come from buffering in network device. When the packets are transmitting through the network, they are queued in a packet buffer. If the queue is stretched to its capacity, the packets are dropped. In case of voice packet, a switch buffer is filling up with both voice packet and background packet. While buffer is growing by inbound packet, the end-to-end delay is increasing. Bolot [9] studies a queuing model to characterize end-to-end packet delay and loss in Internet and proposed waiting time in queue delay of probe packet depends on amount of background traffic. When the buffer is full, packets drop and result in degradation of voice quality.

4 Experiment

4.1 Experimental Setup

To study the effect of background traffic packet size on VoIP's voice quality, we setup the experiment as shown in Fig. 1. Two network switches A and B are the bottleneck node in network and all links use 100 Mbps speed. Voice traffic was generated by RTP Tool Box and PacketGen software [10][11]. Both sets of software are softphone that specializes in initiate and manipulate multiple voice sessions simultaneously. When PacketGen answers the call, it will play the speech with 1.30 minute long to RTP Tool Box and RTP Tool Box will play the same speech back to PacketGen. For the measurement, we chose the G.711, G.726 (32 bit), G.729 and GSM encoding scheme for the voice. The background traffic was generated using Bricks software. Bricks is a simulation tool used to test the bandwidth of a given network. It works by flooding an IP with chunks of data known as "Bricks" and measures the

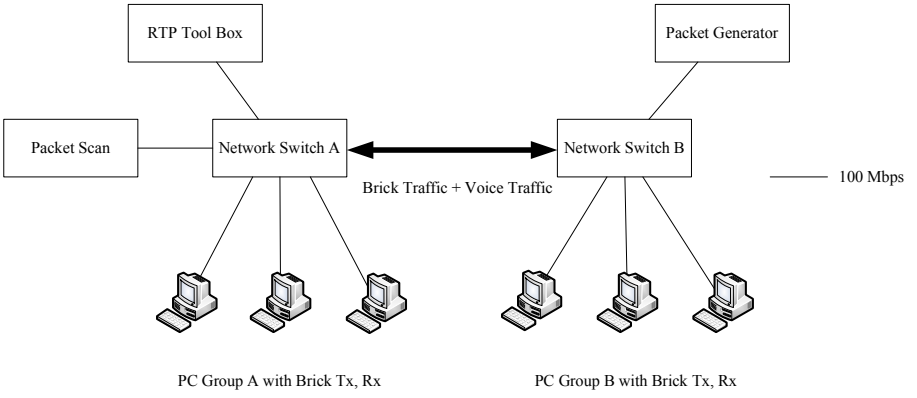


Fig. 1. Experimental diagram

datarate of information being sent. Bricks comes with two programs, one for the client and one for the server. The server will act as the data output while the client will receive the chunks of data. For the measurement, we chose two groups of fifteen PCs to run Bricks. PC in each group will run Bricks Transmitter and Bricks Receiver to transmit and receive packet with each other. We chose 64, 800 and 1500 packet sizes to generated background traffic from Bricks. The three packet sizes represent the "small" "medium" and "big" sizes of TCP packets. The assessment of speech quality was done by PacketScan [12]. PacketScan is a real-time VoIP analyzer that captures live IP traffic, segregates them into SIP/H323 calls and collects statistics about the calls.

4.2 Experimental Results

First, we examine the MOS of speech quality with no background traffic. We use RTP Tool Box to generated 60 SIP calls to Packet Generator and measure the average MOS of 60 calls. The result was shown in the table 2.

Table 2. MOS with no background traffic

Codec	Rate (Kbps)	Packet Size (byte)	MOS	Standard MOS
G.711	64	170	4.20	4.30
G.729	8	30	3.90	3.92
G.726 (32)	32	90	3.95	3.85
GSM	12.2	40	3.48	3.50

Table 2 presents a summary of the MOS from chosen codec. It was found that all MOS was closely to the standard value. Next, we examine the MOS of speech quality with background traffic. We use Bricks software to generate TCP background traffic of 64, 800 and 1500 bytes and measure the average 60 calls over G.711 SIP MOS. The result is shown in Fig. 2.

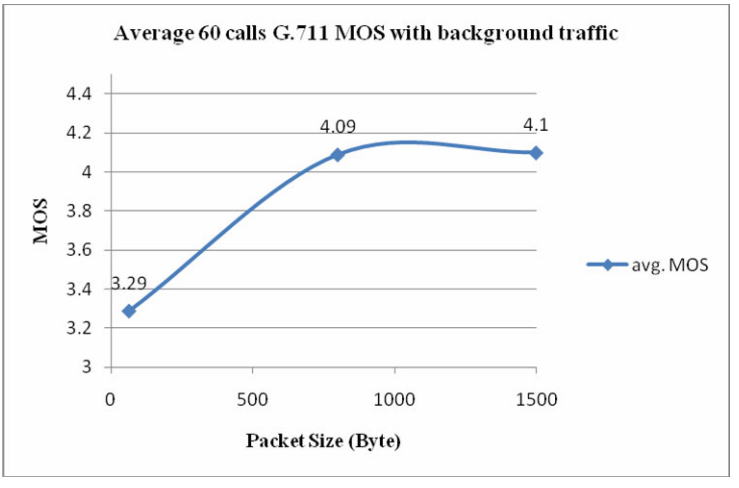


Fig. 2. Average 60 calls G.711 MOS with background traffic

As we have seen from Fig. 2, MOS of voice with 64-byte packet of background traffic was dropped to 3.29 and rise close to standard with the bigger packet size of background traffic. It seems that the small packets size has more effect to voice quality. From the result of the first test, we decide to verified a smaller packet size to the voice quality and add smaller packet size in range of 64-100 byte to the next scenarios with 4 types of codec and get the results as shown in figures 4, 5, 6, and 7.

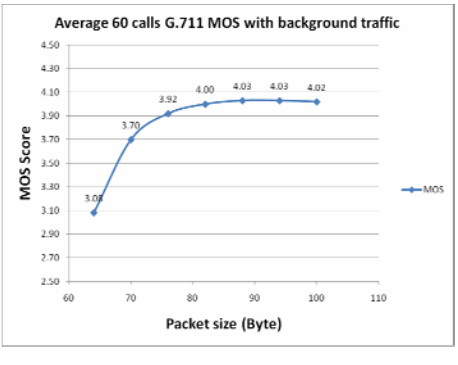


Fig. 3. MOS of the average of 60 calls over G.711 with background traffic

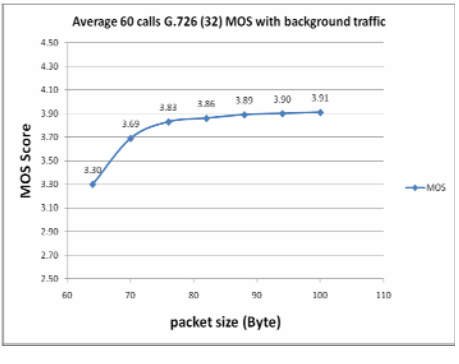


Fig. 4. MOS of the average of 60 calls over G.726 with background traffic

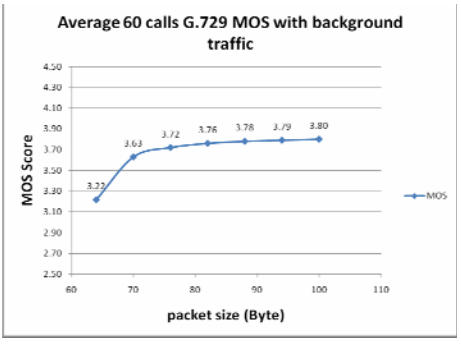


Fig. 5. MOS of the average of 60 calls over G.729 with background traffic

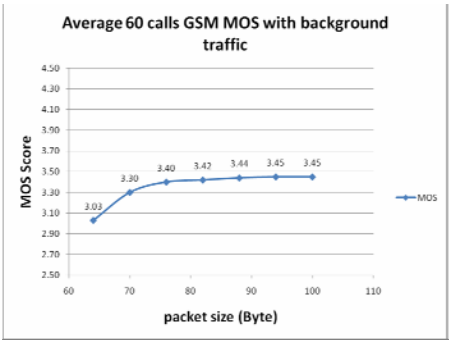


Fig. 6. MOS of the average of 60 calls over GSM with background traffic

5 Discussion

Figures 3 show MOS of the average of 60 calls over G.711 codec with various background traffics. MOS was fall to 3.08 when there is a 64 byte background traffic flow in a network. When background traffic changed to 79 byte, MOS was rise with rapidly to 3.92 and insignificant change after 82 byte background traffic. From the results, it is clear that MOS drops when the network was crowded with small packet traffic. This can be explained with the relation of transmit rate and packet size. In the bottleneck node buffer, a packet is a unit for management irrespective of its length [13]. When you reduce the size of packet, Bricks will increase the number of transmitted packets and the network switch buffer will congest more often. This results in

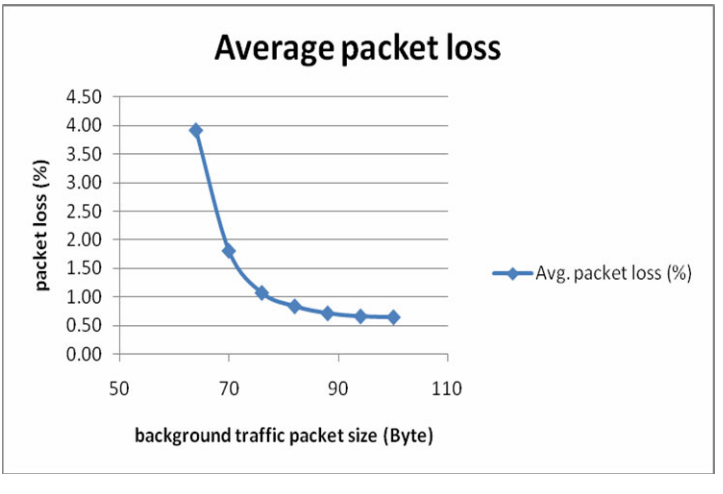


Fig. 7. Average packet loss in each traffic packet size

G.729 and GSM, an effect from small packet traffic seems to less than in G.711 and G.726.the packet loss of VoIP traffic and lead to the degradation of VoIP speech quality. When a network was crowded with a small packet, a codec that produce a smaller packet has a better tolerate to packet loss than a codec with bigger packet. In case of Figure 7 shows average packet loss in each sample codec. At size 64 byte, the average packet losses are 3.91% and reduce rapidly to 1.81% at size 70 byte. In 100-1500 byte case, there are not significant differences in packet losses

6 Conclusion and Future Work

In this paper, we have shown the effect of background traffic packet size with voice quality. It is clear that the packet size of background traffic should be considered in designing or deploying VoIP system. With the data of packet distribution in a network, it can approximately predict the voice quality when deploy VoIP in this network. A packet size distribution will be varied based on real application such as social network, online games, peer-to-peer traffic, etc. The results also shown that optimizing voice packetization under different traffic packet size can improve a voice quality. The optimization of voice packet size with different background traffic can use as a mean for adaption and lead to adaptive packetization VoIP. However, this experimental does not take several important factors into account such as delay, bandwidth, etc. Further research will focus on a detection of packet loss mechanism in VoIP.

References

1. ITU-T P.800: Methods for subjective determination of transmission quality (1996)
2. ETSI Speech Communication Quality for Mouth to Ear for 3.1 kHz Handset Telephony across Networks. Technical Report ETR250 (1996)
3. ITU-T Recommendation G.107: The Emodel, a computational model for use in transmission planning (December 1998)
4. ITU-T Recommendation G.108: Application of the Emodel: a planning guide (1998)
5. Cox, R., Perkins, M.: Results of a subjective listening test for G.711 with frame erasure concealment, AT&T contribution to T1A1.7/99-016 (1999)
6. Rosenberg, J.: G.729 Error Recovery for Internet Telephony, Project Report, Columbia University (1997)
7. Koodli, R., Ravikanth, R.: One-way Loss Pattern Sample Metrics <draft-ietf-ippm-loss pattern-03.txt>, Internet Draft, Internet Engineering Task Force (2002)
8. Sun, L., Wade, G., Lines, B., Ifeachor, E.: Impact of Packet Loss Location on Perceived Speech Quality. In: Proceedings of 2nd IP-Telephony Workshop (IPTEL 2001), pp. 114–122. Columbia University, New York (2001)
9. Bolot, J.-C.: Characterizing End-to-end Packet Delay and Loss in the Internet. Journal of High Speed Networks 2(3), 305–323 (1993)
10. GL Communications Inc. RTP Toolbox product, `http://www.gl.com/rtptoolbox.html`

11. GL Communications Inc. PacketGen product,
http://www.gl.com/packetgen.html
12. GL Communications Inc. PacketScan product,
http://www.gl.com/packetscan.html
13. Sawashima, H., Sunahara, Y.H.H.: Characteristics of UDP packet loss: Effect of tcp traffic. In: Proceedings of INET 1997: The Seventh Annual Conference of the Internet Society, Kuala Lumpur (1997)

Classification of Internal Carotid Artery Doppler Signals Using Hidden Markov Model and Wavelet Transform with Entropy

Harun Uğuz and Halife Kodaz

Selçuk University, Department of Computer Engineering, 42075, Konya, Turkey
{harun_uguz,hkodaz}@selcuk.edu.tr

Abstract. Doppler ultrasound has been usually preferred for investigation of the artery conditions in the last two decade, since it is a non-invasive method which is not risky. In this study, a biomedical system based on Discrete Hidden Markov Model (DHMM) has been developed in order to classify the internal carotid artery Doppler signals recorded from 191 subjects (136 of them had suffered from internal carotid artery stenosis and rest of them had been healthy subjects). Developed system comprises of three stages. In the first stage, for feature extraction, obtained Doppler signals were separated to its sub-bands using Discrete Wavelet Transform (DWT). In the second stage, entropy of each sub-band was calculated using Shannon entropy algorithm to reduce the dimensionality of the feature vectors via DWT. In the third stage, the reduced features of carotid artery Doppler signals were used as input patterns of the DHMM classifier. Our proposed method reached 97.38% classification accuracy with 5 fold cross validation (CV) technique. The classification results showed that purposed method is effective for classification of internal carotid artery Doppler signals.

Keywords: Discrete Hidden Markov model; Doppler signal; Carotid artery; Wavelet Transform; Entropy.

1 Introduction

Carotid artery is a disease negatively affecting the vessels going to the head and brain. The symptoms of carotid artery disease are stenosis and occlusions occurring in the internal carotid artery. These symptoms are led by the internal carotid artery plaques [1]. Angiography and blood tests are the methods which can be referred to for diagnostics of the disease. On the other hand, angiography and blood tests are not preferred since they are invasive. Doppler technique is non-invasive, portable, easy to apply, reliable and tolerable technique as well as being cheaper than these invasive methods [2]. Therefore this technique is usually used for demonstrates the flow characteristic of carotid arteries.

In the literature, there are various studies on classification of the carotid artery Doppler signals. Özşen et al. extracted the features with AR method and classified these features using a new Artificial Immune Systems classifier [3]. Ceylan et al. used

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 183–191, 2010.

Complex Valued Artificial Neural Network (CVANN) structure to classify carotid artery Doppler signals using Principal Component Analysis and Fuzzy c-means Clustering (FCM) as feature extraction methods before the CVANN classifier [4]. Özbay & Ceylan used Fast Fourier Transform, Hilbert Transform, and Welch Method with different window types. They investigated effects of window types on classification of carotid artery Doppler signals [5]. Polat et al. used Support Vector Machine (SVM) with a fuzzy-weighted preprocessing step in classification progress [6]. In Güler and Übeyli's studies, attributes attained from the carotid artery Doppler signals by the help of the various spectral analysis methods were classified according to neural network-based classification systems [7-9].

In this study, a biomedical-based application has been developed for the classification of internal carotid artery Doppler signals recorded from 191 subjects. Application is mainly comprised of three stages, namely as being feature extraction, dimension reduction, and classification. At the feature extraction stage, obtained Doppler signals were separated to its sub-bands using Discrete Wavelet Transform (DWT). But, for the classifier a large number of input parameters will increase computationally intensive and time consuming. Therefore choosing fewer features to represent data set was aimed rather than a large number of features obtained with DWT. In this way, at dimension reduction stage, entropy of each sub-band was calculated using Shannon entropy algorithm to reduce the dimensionality of the feature vectors via DWT. At classification stage, the reduced features internal carotid artery Doppler signals were used as input patterns of the Discrete Hidden Markov Model (DHMM) classifier. Classification success at a rate by 97.38% was achieved in purposed method. Classification results were displayed the fact that, purposed method was effective in the classification of internal carotid artery Doppler signals.

2 Materials and Methods

Fig.1. shows the procedure used in the development of the classification system. It consist of five parts: (a) measurement of internal carotid artery Doppler signals, (b) feature extraction of internal carotid artery Doppler signals by using DWT, (c) dimension reduction with Shannon entropy algorithm (d) classify internal carotid artery Doppler signals with DHMM, (e) classification results (as healthy or stenosis).

2.1 Measurement of Internal Carotid Artery Doppler Signals

Doppler signals recorded from 136 patients as well as of 55 healthy people in Radiology clinics of Medical Center of Firat University. Internal carotid artery examinations were performed with a Doppler unit using a 6.2 -8.4 MHz linear transducer has been used. Various tools consist of Doppler unit (Toshiba SSA-770 Aplio 80, Toshiba, Tokyo, Japan), an analog/digital interface board and Personal Computer (PC) were used for raw data obtaining. The analog Doppler unit works in not only continuous mode but also pulse wave mode. The probe was most often placed at an angle of 60° towards the internal carotid artery.

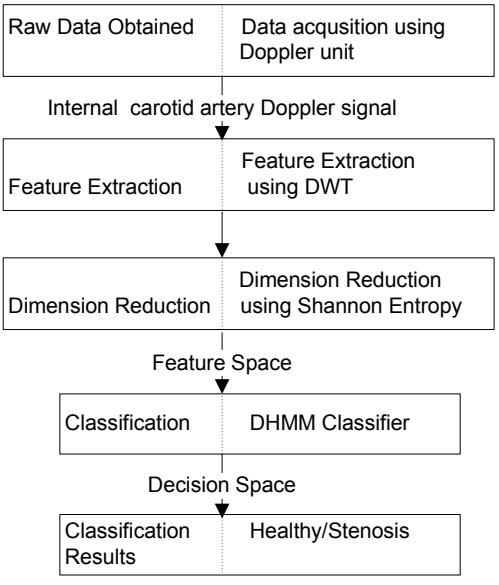


Fig. 1. The structure of the suggested classification system

2.2 Wavelet Transform

Wavelet is a time frequency analysis tool which effectively analysis non-stationary signals [10]. Wavelets are such mathematical functions that allow us to work on each signal component by decomposing signals into their components residing in different frequencies. Main aspect of Wavelet transform is formed the mother wavelet function which will give useful and sufficient information about the observed signal. It has advantages on observing non-stationary signals and signals which can have sharp peak values over traditional Fourier methods. In contrast to traditional Fourier transform, Wavelet transform gives short time intervals for the high-frequency bands and long time intervals for the low-frequency bands [11]. Therefore, wavelet transform gives precise frequency information at low frequency and precise time information at high frequency [12]. As a matter of fact, using wavelet transform is appropriate since it gives the information about the signal both in frequency and time domain [13]. Details for mastering the art of wavelet transform are given in [14].

2.3 Entropy

Entropy measurement is an ideal method in order to measure the level of disorder of a non-stationary signal [15]. Besides, entropy is also being used for the purpose of measuring average amount of information that an event contains [16]. An Entropy-based criterion describes information-related properties for an accurate representation of a given signal [17]. Entropy is a common concept in many fields, mainly in signal processing. Studies that have been reported in the literature, entropy concept has been successfully applied to signal processing [17-21]. For signal processing, Shannon, Norm, Threshold and Logarithmic Energy are the most common entropy calculation

methods. Shannon entropy which has been proved to work well in signal processing in literature [18, 21] was used in this study.

2.4 Discrete Hidden Markov Model

A DHMM is a stochastic Finite State Machine (FSM). A DHMM is a double-layered finite state stochastic process, with a hidden Markovian process that controls the selection of the states of an observable process. In general a DHMM has *N* states, and transitions are available among the states. At different times, the system is in one of these states; each transition between the states has an associated probability, and each state has an associated observation output (symbol). A DHMM is characterized by the followings [22]:

1. N, the number of states of the model. The set of individual states is denoted as $S=\{S_1, S_2, \ldots, S_N\}$, and state at time t as q_t .
2. T, the number of observations. A typical observation sequence is denoted as

$$O=\{O_1, O_2, \ldots, O_T\}. \tag{1}$$

3. $A=\{a_{ij}\}$, the state transition probability distribution, of size *NxN* , defines the probability of transition from state *i* at time *t*, to state *j* at time *t+1*.

$$a_{ij}=P(q_{t+1}=S_j|q_t=S_i),\ 1\leq i,j \leq N \tag{2}$$

4. $B=\{b_j(k)\}$, (i=1,…,N;k=1,…,M) probability distribution of observation symbols for each state. b_j is observation symbol probability in state *j.* M is the number of the observation symbols.
5. The initial state distribution, $\pi =\{ \pi_i \}$, defines the probability of any given state, where

$$\pi_i =P(q_1=S_j),\ 1\leq i \leq N \tag{3}$$

A complete specification of a DHMM requires specification of model parameters, specification of the three probability measures A, B, π. DHMM parameters use the following set:

$$\lambda =\{ A, B, \pi \} \tag{4}$$

3 Experimental Results

3.1 Feature Extraction Using DWT

In this study, DWT has been used in order to extract feature vectors from the internal carotid artery Doppler signals. Proper wavelet selection and determining decomposition level number plays a significant role in analysis by using wavelet transform method. This number of decomposition levels is figured out based on the dominant frequency components of the signal. The levels are determined so that those parts of the signal that has well correlation with the required frequencies for classification of the signal and that are retained in the wavelet coefficients [7, 23, 24]. In this study,

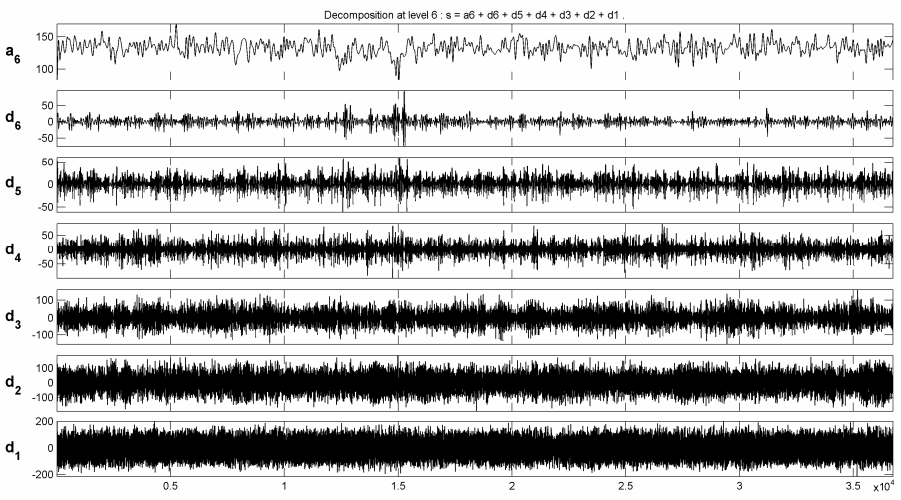


Fig. 2. Wavelet detail coefficients at the first-six decomposition levels (d_1–d_6) and wavelet approximation coefficient at six decomposition level (a_6) for a healthy subject

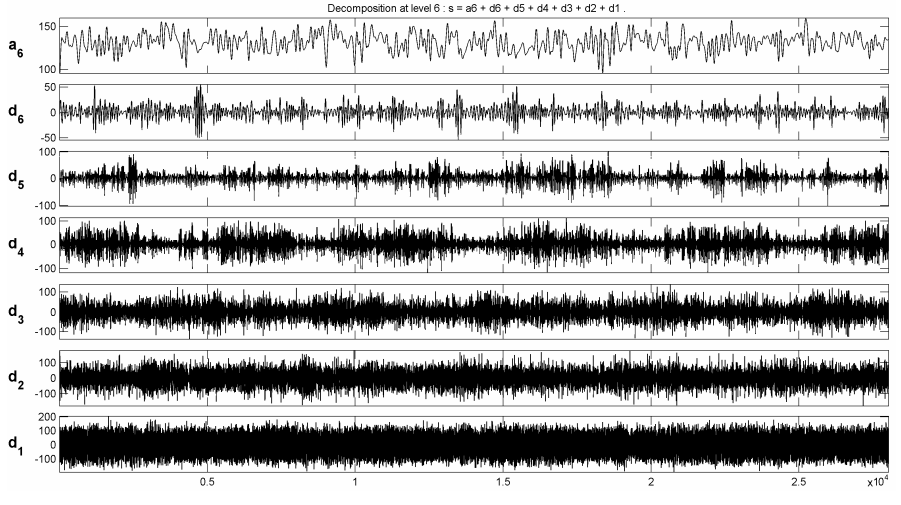


Fig. 3. Wavelet detail coefficients at the first-six decomposition levels (d1-d6) and wavelet approximation coefficient at the six decomposition level (a6) for an unhealthy subject suffering from internal carotid artery stenosis

it has been determined empirically the number of decomposition levels as 6. Thus situated, internal carotid artery Doppler signals are separated into (d_1-d_2-d_3-d_4-d_5-d_6) detailed sub-bands and lastly into approximation sub-band a_6. Since accuracy of the classification is closely related with the selected wavelet type for the application, it is important to figure out the wavelet type to be chosen. Usually, wavelet type selection is done after various tests are performed with different wavelet types when the one

which gives the maximum efficiency for the concerning application is figured out. Both in temporal and frequency domains, Daubechies wavelet has good localizing properties. Order 6 (*db*6) wavelet of Daubechies lead more suitable detection of changes of the signals under study.

Fig. 2 and 3 show wavelet detail coefficients at the first six decomposition levels, and wavelet approximation coefficients at the six level for a healthy subject, and unhealthy subject suffering from internal carotid artery stenosis respectively. As seen in Fig. 2 and 3, there are apparent differences between the graphics of the internal carotid artery Doppler signal of a healthy and unhealthy subject. Therefore, such a classification system, being established by taking such variances in DWT graphics into consideration, enables for deciding on respective disease.

3.2 Application of Entropy to Wavelet Transformed Doppler Signals

Feature extraction stage significantly affects the classification performance. Instead of making complicated classifier structures, extracting better features amongst the patterns is the main aspect of recent pattern recognition systems. At the same time for the classifier a large number of input parameters will increase computationally intensive and time consuming. Besides, presence of excess number of redundant, irrelevant, and noisy input variables may hide the meaningful variables in the data set. Therefore choosing fewer features to represent data set was aimed rather than a large number of features.

At feature extraction stage of this study, as the first step, the wavelet coefficients corresponding to the d_1-d_6 and a_6 frequency bands of the Doppler signals were computed. Here, it is required to reduce the feature dimensions of the feature vectors due to excessive number of high dimensional feature vectors faced. In this point, Shannon entropy has been used in order to reduce the dimensionality of the feature vectors. The values of Shannon entropy of each sub-band of Doppler signals were calculated by equation 5:

$$E(s) = -\sum_i s_i^2 \log(s_i^2) \tag{5}$$

where s is the signal and s_i are the coefficients of s in an orthonormal basis [25].

In this way, the feature vector was extracted by calculating the 7 wavelet entropy (six detailed and one approximation) values for per internal carotid artery Doppler signal. These wavelet entropy values were obtained via trial and error to give the best classification performance. Thus, the feature vector had a lower dimension that includes most of the useful information from the original vector.

3.3 Classification of Internal Carotid Artery Doppler Signals Using DHMM

From the dimension reduction stage via Shannon entropy method, it was proceeded to the classification stage of Doppler signals. 7 features, obtained after the application of Shannon entropy method for each sub-band were applied to DHMM classifier as input. In this study, 5-fold cross validation technique was used.

For a DHMM classifier, the Vector quantization (VQ) process and the number of states are the most important factors which affect the classification performance.

Therefore, we made experiments by changing the VQ dimension and the number of states in order to find the best classification performance. Fuzzy-C-Means algorithm was used for VQ process in this study. The architecture of the DHMM is shown in Table 1.

Table 1. DHMM architecture

Modeling Description	Settings
The best state numbers	9
The best VQ size	9
Evaluating problem solved	Forward-Backward
Decoding problem solved	Forward Algorithm
Training problem solved	Baum-Welch
Maximum training iterations	5000
Type	Ergodic Model

The confusion matrix including DHMM classification results are given in Table 2. Confusion matrix was created taking into account the test errors in all folds (5 folds). According to the confusion matrix in Table 3, 3 subjects with stenosis were incorrectly classified as healthy, and 2 healthy subjects were incorrectly classified as patients with stenosis.

Table 2. Confusion matrix for the DHMM classification results

Output/Desired	Healthy	Stenosis
Healthy	51	3
Stenosis	2	134

Table 3. Statistical parameters of the classifier method

Classifier	Specificity	Sensitivity	Accuracy
DHMM	94.55%	98.53%	97.38%

The classification performance of DHMM based method was evaluated using the statistical parameters like sensitivity and specificity. Table 3 presents result depending on sensitivity/specificity rate. The classification results showed that purposed method is effective for classification of internal carotid artery Doppler signals.

4 Conclusions

In this study, a biomedical system based on DHMM has been developed in order to classify the internal carotid artery Doppler signals. Developed system comprises of three stages. In the first stage, for feature extraction, obtained internal carotid artery Doppler signals were separated to its sub-bands using discrete wavelet transform (DWT). In the second stage, entropy of each sub-band was calculated using Shannon

entropy algorithm to reduce the dimensionality of the feature vectors via DWT. The reduced features of internal carotid artery Doppler signals were used as input patterns of the DHMM classifier. In experiments, DHMM based method reached 97.38% classification accuracy with 5 fold CV technique. The classification results showed that purposed method is effective for classification of internal carotid artery Doppler signals. Above all, development of this kind of decision-support systems will provide assistance to physicians with lacking experience, and skill in diagnosing the internal carotid artery disease, by simplifying this diagnosis process.

Acknowledgement

This study has been supported by Scientific Research Project of Selçuk University.

References

1. Baker, W.H.: Diagnosis and Treatment of Carotid Artery Disease. Futura Publishing Company Inc., New York (1985)
2. Miranda, P., Lagares, A., Alen, J., Perez-Nunes, A., Arrese, I., Lobato, R.D.: Early transcranial Doppler after subarachnoid hemorrhage: clinical and radiological correlations. Surgical Neurology 65(3), 247–252 (2006)
3. Özşen, S., Kara, S., Latifoğlu, F., Güneş, S.: A new supervised classification algorithm in artificial immune systems with its application to carotid artery Doppler signals to diagnose atherosclerosis. Computer Methods and Programs in Biomedicine 88, 246–255 (2007)
4. Ceylan, M., Ceylan, R., Dirgenali, F., Özbay, Y.: Classification of carotid artery Doppler signals in the early phase of atherosclerosis using complex-valued artificial neural network. Computers in Biology and Medicine 37(1), 28–36 (2007)
5. Özbay, Y., Ceylan, M.: Effects of window types on classification of carotid artery Doppler signals in the early phase of atherosclerosis using complex-valued artificial neural network. Computers in Biology and Medicine 37(3), 287–295 (2007)
6. Polat, K., Kara, S., Latifoğlu, F., Güneş, S.: Pattern detection of atherosclerosis from carotid artery Doppler signals using fuzzy weighted pre-processing and least square support vector machine (LSSVM). Ann. Biomed. Eng. 35(5), 724–732 (2007)
7. Güler, İ., Übeyli, E.D.: A recurrent neural network classifier for Doppler ultrasound blood flow signals. Pattern Recognition Letters 27(1), 1560–1571 (2006)
8. Güler, İ., Übeyli, E.D.: Implementing wavelet/probabilistic neural networks for Doppler ultrasound blood flow signals. Expert Systems with Applications 33(1), 162–170 (2007)
9. Übeyli, E.D., Güler, İ.: Neural network analysis of internal carotid arterial Doppler signals: predictions of stenosis and occlusion. Expert Systems with Applications 25, 1–13 (2003)
10. Subasi, A.: EEG signal classification using wavelet feature extraction and a mixture of expert model. Expert Systems with Applications 32(4), 1084–1093 (2007)
11. Wu, J.D., Kuo, J.M.: An automotive generator fault diagnosis system using discrete wavelet transform and artificial neural network. Expert Systems with Applications 36(6), 9776–9783 (2009)
12. Ocak, H.: Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy. Expert Systems with Applications 36, 2027–2036 (2009)

13. Purushotham, V., Narayanan, S., Suryanarayana, A.N.P.: Multi-fault diagnosis of rolling bearing elements using wavelet analysis and hidden Markov model based fault recognition. NDT&E International 38(8), 654–664 (2005)
14. Mallat, S.G.: A theory for multiresolution signal decomposition: The wavelet representation. IEEE Transactions on Pattern Anal Machine Intelligence 11(7), 674–693 (1989)
15. Tong, S., Bezerianos, A., Paul, J., Zhu, Y., Thakor, N.: Nonextensive entropy measure of EEG following brain injury from cardiac arrest. Physica A: Statistical Mechanics and its Applications 305(3-4), 619–628 (2002)
16. Principe, J.C., Euliano, N.R., Lefebvre, W.C.: Neural and Adaptive Systems. John Wiley & Sons, New York (2000)
17. Turkoglu, I., Arslan, A., Ilkay, E.: An expert system for diagnosis of the heart valve diseases. Expert Systems with Applications 23, 229–236 (2002)
18. Turkoglu, I., Arslan, A., Ilkay, E.: An intelligent system for diagnosis of the heart valve diseases with wavelet packet neural Networks. Computers in Biology and Medicine 33, 319–331 (2003)
19. Zhang, X.S., Roy, R.J.: Derived fuzzy knowledge model for estimating the depth of anesthesia. IEEE Transactions on Biomedical Engineering 48(3), 312–323 (2001)
20. Kannathal, N., Choo, M.L., Acharya, U.R., Sadasivan, P.K.: Entropies for detection of epilepsy in EEG. Computer Methods and programs in Biomedicine 80, 187–194 (2005)
21. Yildiz, A., Akin, M., Poyraz, M., Kirbas, G.: Application of adaptive neuro-fuzzy inference system for vigilance level estimation by using wavelet-entropy feature extraction. Expert Systems with Applications 36, 7390–7399 (2009)
22. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE 77(2), 257–286 (1989)
23. Kandaswamy, A., Kumar, C.S., Ramanathan, R.P., Jayaraman, S., Malmurugan, N.: Neural classification of lung sounds using wavelet coefficients. Computers in Biology and Medicine 34(6), 523–537 (2004)
24. Kiymik, M.K., Akin, M., Subasi, A.: Automatic recognition of alertness level by using wavelet transform and artificial neural network. Journal of Neuroscience Methods 139, 231–240 (2004)
25. Yildiz, A., Akin, M., Poyraz, M., Kirbas, G.: Application of adaptive neuro-fuzzy inference system for vigilance level estimation by using wavelet-entropy feature extraction. Expert Systems with Applications 36, 7390–7399 (2009)

Genetic Algorithm with Species for Regularization Network Metalearning

Roman Neruda and Petra Vidnerová

Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 18207 Prague 8, Czech Republic
roman@cs.cas.cz

Abstract. Regularization networks are one of the important methods for supervised learning. They benefit from very good theoretical background, though their drawback is the presence of metaparameters. The metaparameters are typically supposed to be given by an user. In this paper, we develop a method for finding optimal values for metaparameters, namely type of kernel function, kernel's parameter and regularization parameter. The method is based on genetic algorithms with different species for different kinds of kernels. The method is demonstrated on experiments.

Keywords: Regularization, Neural networks, Metalearning, Genetic algorithms.

1 Introduction

Regularization theory presents a sound framework to solving supervised learning problems. Regularization networks (RN) benefit from very good theoretical background [1,2,3] and a simple and effective learning algorithm [4]. Their disadvantage is a presence of metaparameters that are supposed to be known in advance. These metaparameters are a type of kernel function and a regularization parameter. In addition, the kernel function typically has additional parameter, for instance Gaussian kernel has a width.

In this paper we introduce a method for optimization of RN metaparameters. The method is based on minimization of cross-validation error, which is an estimate of generalization ability of a network, by means of genetic algorithms. Different species are employed corresponding to different kinds of kernel functions. The paper is organized as follows. In the next section we introduce the regularization network. In Section 3 the cross-validation error is briefly introduced. Section 4 describes the genetic parameter search. Section 5 contains results of our experiments. Conclusion can be found in Section 6.

2 Regularization Networks

To develop regularization networks we formalize the problem of supervised learning as an function approximation problem. We are given a set of examples

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 192–201, 2010.

$\{(\boldsymbol{x}_i, y_i) \in R^d \times R\}_{i=1}^N$ obtained by random sampling of some real function f and we would like to find this function. Since this problem is ill-posed, we have to add some a priori knowledge about the function. We usually assume that the function is *smooth*, in the sense that two similar inputs corresponds to two similar outputs and the function does not oscillate too much. This is the main idea of the regularization theory, where the solution is found by minimizing the functional (1) containing both the data and smoothness information.

$$H[f] = \frac{1}{N}\sum_{i=1}^{N}(f(\boldsymbol{x}_i) - y_i)^2 + \gamma\Phi[f], \tag{1}$$

where Φ is called a *stabilizer* and $\gamma > 0$ is *the regularization parameter* controlling the trade-off between the closeness to data and the smoothness of the solution. The regularization scheme (1) was first introduced by Tikhonov [5] and therefore it is called a Tikhonov regularization. The regularization approach has good theoretical background, it was shown that for a wide class of stabilizers the solution has a form of feed-forward neural network with one hidden layer, called *regularization network*, and that different types of stabilizers lead to different types of regularization networks [3,4].

Poggio and Smale in [4] proposed a learning algorithm (Alg. 1) derived from the regularization scheme (1). They choose the hypothesis space as a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_K$ defined by an explicitly chosen, symmetric, positive-definite kernel function $K_{\boldsymbol{x}}(\boldsymbol{x}') = K(\boldsymbol{x}, \boldsymbol{x}')$. The stabilizer is defined by means of norm in $\mathcal{H}_K$, so the problem is formulated as follows:

$$\min_{f\in\mathcal{H}_K} H[f], \text{where } H[f] = \frac{1}{N}\sum_{i=1}^{N}(f(\boldsymbol{x}_i) - y_i)^2 + \gamma||f||_K^2. \tag{2}$$

The solution of minimization (2) is unique and has the form

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} w_i K_{\boldsymbol{x}_i}(\boldsymbol{x}), \qquad (N\gamma I + K)\boldsymbol{w} = \boldsymbol{y}, \tag{3}$$

where I is the identity matrix, K is the matrix $K_{i,j} = K(\boldsymbol{x_i}, \boldsymbol{x_j})$, and $\boldsymbol{y} = (y_1, \ldots, y_N)$.

The solution (3) can be represented by a neural network with one hidden layer and output linear layer. The most commonly used kernel function is Gaussian $K(\boldsymbol{x}, \boldsymbol{x}') = e^{-\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|}{b}\right)^2}$.

The power of the Algorithm 1 is in its simplicity and effectiveness. However, there are the parameters γ and the type of kernel function, and usually another parameter of the kernel function, in case of Gaussian a width. These parameters are supposed to be fixed. We call them *metaparameters*. Then, the algorithm reduces to the problem of solving linear system of equations (4). Since the system has N variables, N equations, K is positive-definite and $(N\gamma I + K)$ is strictly positive, it is well-posed, i.e. is has a unique solution and the solution exists.

Input: Data set $\{\boldsymbol{x}_i, y_i\}_{i=1}^N \subseteq X \times Y$
Output: Function f.

1. Choose a symmetric, positive-definite function $K_{\boldsymbol{x}}(\boldsymbol{x}')$, continuous on $X \times X$.
2. Create $f : X \rightarrow Y$ as $f(\boldsymbol{x}) = \sum_{i=1}^N c_i K_{\boldsymbol{x}_i}(\boldsymbol{x})$ and compute $\boldsymbol{w} = (w_1, \ldots, w_N)$ by solving

$$(N\gamma I + K)\boldsymbol{w} = \boldsymbol{y}, \tag{4}$$

where I is the identity matrix, $K_{i,j} = K(\boldsymbol{x_i}, \boldsymbol{x_j})$, and $\boldsymbol{y} = (y_1, \ldots, y_N)$, $\gamma > 0$.

Algorithm 1. RN learning algorithm

In addition, we need the system to be well-conditioned, i.e. insensitive to small perturbations of the data. That holds for small condition number of the matrix $(N\gamma I + K)$, that is for large $N\gamma$. Note that we are not entirely free to choose γ, because with too large γ we loose the closeness to data. The real performance of the algorithm depends significantly on the choice of metaparameters γ and kernel function. However, their optimal choice depends on a particular data set and there is no general heuristics for setting them.

3 Optimization of Metaparameters

The optimal RN should not only approximate the data from training set, but also posses a good generalization ability, i.e. give good replies also for data points not given in the training set. Our estimate of a generalization ability is a *cross-validation error*. Then we search for such metaparameters that optimize the cross-validation error. In particular, we use a *k-fold cross-validation* that uses partitioning of the original data set S into k subsets $S_1, \ldots, S_k$, so that $\bigcup_i S_i = S$ and $\forall\ i \neq j : S_i \bigcap S_j = \emptyset$. The cross-validation process consists of k trials. In each trial a single subset is retained as the validation set, and the remaining $k-1$ subsets are used as training set, so that each of the k subsamples used exactly once as the validation set. The k results from the folds then can be averaged to produce a single estimation E_{kfolds}:

$$E_{kfolds} = \frac{1}{k} \sum_{i=1}^{k} E(f^{\bigcup_{i \neq j} S_j}, S_i), \tag{5}$$

where f^S is a RN learned on data S and $E(f, S)$ is an error of network f on data S.

Now, let $f^S_{\gamma,K}$ be the RN found by the Alg. 1 with the regularization parameter γ and kernel function K on data set S. Then by $E_{cross}(\gamma, K, S)$ we denote the cross-validation error

$$E_{cross}(\gamma, K, S) = \frac{1}{k}\sum_{i=1}^{k} E(f_{\gamma,K}^{\bigcup_{i\neq j} S_j}, S_i). \quad (6)$$

. We will search for meta-parameters that minimize the cross-validation error (6). It means we will choose a solution f_{γ^*,K^*}^{S} such that

$$[\gamma^*, K^*] = \text{argmin}_{\gamma,d} E_{cross}(\gamma, K, S). \quad (7)$$

4 Genetic Parameter Search

Genetic algorithms (GA) [6] represent a search technique used to find approximate solutions to optimization and search problems. They belong to evolutionary algorithms that use techniques inspired by evolutionary biology such as mutation, selection, and crossover. Genetic algorithms typically work with a population of *individuals* representing abstract representations of feasible solutions. Each individual is assigned a *fitness* that is a measure of how good solution it represents. The better the solution, the higher the fitness value. The population evolves towards better solutions. The evolution starts from a population of completely random individuals and iterates in generations. In each generation, the fitness of each individual is evaluated. Individuals are stochastically selected from the current population (based on their fitness), and modified by means of operators *mutation* and *crossover* to form a new population. The new population is then used in the next iteration of the algorithm. We work with individuals coding the parameters of RN learning algorithm (Alg. 1). They are the type of kernel function, its additional parameters, and the regularization parameter, see figure 1. When the type of the kernel function is known in advance, the individual consists only of the kernel's parameter (i.e. the width in the case of Gaussian kernel) and the regularization parameter. New generations of individuals are created using operators *selection*, *crossover* and *mutation*. Mutation introduces small random perturbation to existing individuals, thus for an individual $I_t = (x_0^t, \ldots, x_q^t) \in P_t$, the I_{t+1} is created as: $x_k^{t+1} = x_k^t + z_k$, where $k = 0, \ldots, q$ and z_k is a random number from normal distribution $N(0, 1)$. The crossover (Fig. 2), creates two new individuals from two existing individuals by choosing new parameters values randomly in the interval formed by the old values. Thus, $I_t = (x_0^t, \ldots, x_q^t) \in P_t$ and $I'_t = (y_0^t, \ldots, y_q^t) \in P_t$ are crossed-over as: $x_k^{t+1} = v_k x_k^t + (1 - v_k) y_k^t$, where $k = 0, \ldots, q$ and v_k is a random number from uniform distribution $R(0, 1)$ (and the same holds for y_k^{t+1}).

To work with individuals representing different kernel types we introduce species. Individuals with different kernel functions represent different species, each specie forms one subpopulation. The selection is performed on the whole population and the selected individual is inserted into subpopulation according its kernel type. Crossover is then performed only among the individuals of same

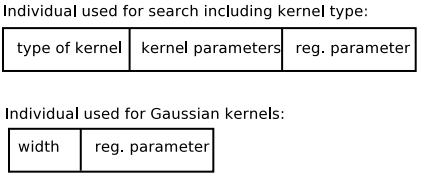


Fig. 1. Illustration of main features of the genetic search algorithm 2: Encoded individuals of the algorithm

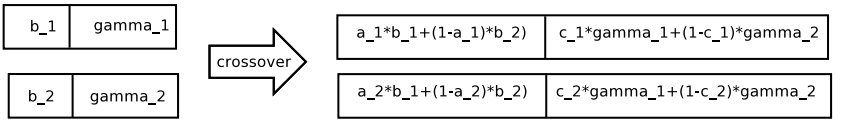


Fig. 2. Illustration of main features of the genetic search algorithm 2: Arithmetic crossover for Gaussian kernels

Input: Data set $S = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N} \subseteq \Re^n \times \Re$
Output: Parameters γ and K.

1. Create randomly an initial population $P^0 = P^0_{K_1} \cup \ldots \cup P^0_{K_n}$, where K_i is a particular kernel function and each $P^0_{K_i}$ has M individuals $I^0_{K_i,1} \ldots I^0_{K_i,M}$. Each $I^0_{K_i,j} = \{K_i, p_j, \gamma_j\}$, where p_j is number of units, and γ_j is a regularization parameter.
2. $i \leftarrow 0$
3. $\forall j : P^{i+1}_{K_j} \leftarrow$ empty set
4. for $j = 0$ to $n * M$:
   (a) $I \leftarrow selection(P^i)$
   (b) insert I into P^{i+1}_K such that $I = \{K, p, \gamma\}$
5. for $j = 0$ to n:
   (a) for $k = 1$ to $\frac{|P^{i+1}_{K_j}|}{2}$:
   with probability p_{cross}: $(I^{i+1}_{K_j,2*k}, I^{i+1}_{K_j,2*k+1}) \leftarrow crossover(I^{i+1}_{K_j,2*k}, I^{i+1}_{K_j,2*k+1})$
   (b) for $k = 1$ to $|P^{i+1}_{K_j}|$:
   with probability p_{mutate}: $I^{i+1}_{K_j,k} \leftarrow mutate(I^{i+1}_{K_j,k})$
6. $\forall I \in P^{i+1}$:
   (a) $E \leftarrow E_{cross}(\gamma, K, p, S)$, where $I = \{K, p, \gamma\}$
   (b) fitness(I) $\leftarrow$ C - E
7. $P^{i+1} \leftarrow P^{i+1}_{K_1} \cup \ldots \cup P^{i+1}_{K_n}$
8. $i \leftarrow i + 1$
9. goto 3 and iterate until the fitness stops increasing

Algorithm 2. Genetic parameter search

subpopulation. Since we want to minimize the cross-validation error, the fitness should reflect it. So the lower the cross-validation error is, the higher the fitness value is. See Algorithm 2 for the sketch of the algorithm.

5 Experiments

For our experiments we used the collection of benchmark problems *Proben1* [7]. This collection has been proposed as a reference set of several classification and approximation problems with results gathered for large class of methods. Thus, it can be used to compare the performance of new algorithms with respect to standard models, such as back propagation and others. The tasks are listed in Table 1. Each task is present in three variants corresponding to three different partitioning to training and testing sets.

The following procedure is used for experiments:

1. find the values for γ and K using genetic parameter search
2. use the whole training set and the parameters found by Step 1 to estimate the weights of RN
3. evaluate error on the testing set

The error is computed as:

$$E = 100\frac{1}{Nm}\sum_{i=1}^{N}||\boldsymbol{y}_i - f(\boldsymbol{x}_i)||^2, \tag{8}$$

where $||\cdot||$ denotes the Euclidean norm.

The standard numerical library LAPACK [8] was used to solve linear systems.

The population had 4 subpopulation, corresponding to kernel functions: Gaussian, Multi-quadratic, Inverse Multi-quadratic, and Sigmoid (see Fig. 3 and Tab. 2).

Table 1. Overview of Proben1 tasks. Number of inputs (n), number of outputs (m), number of samples in training and testing sets (N_{train},N_{test}). Type of task: approximation or classification.

Task name	n	m	N_{train}	N_{test}	**Type**
cancer	9	2	525	174	class
card	51	2	518	172	class
flare	24	3	800	266	approx
glass	9	6	161	53	class
heartac	35	1	228	75	approx
hearta	35	1	690	230	approx
heartc	35	2	228	75	class
heart	35	2	690	230	class
horse	58	3	273	91	class

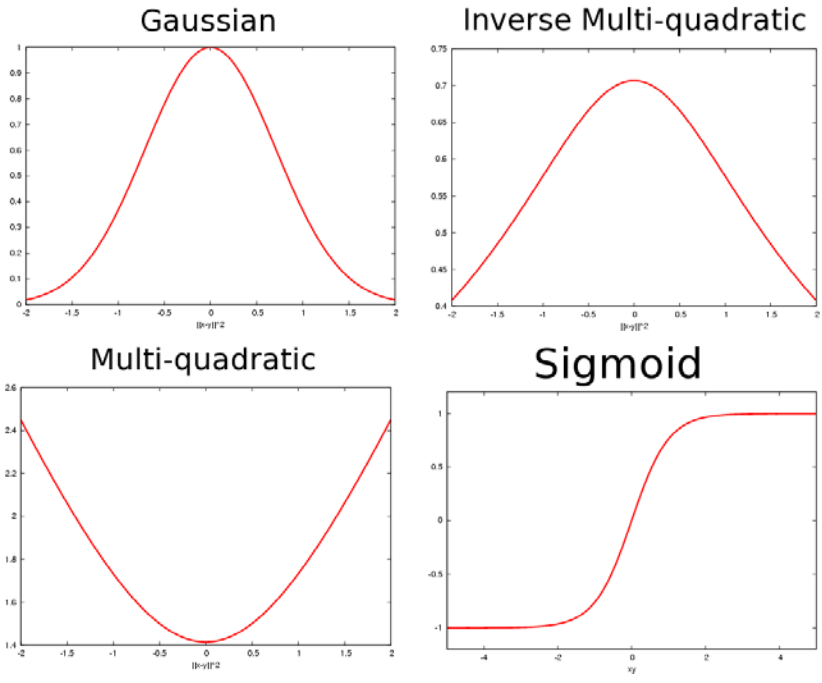


Fig. 3. Kernel functions

Table 2. Kernel functions

Gaussian	$K(x,y) = e^{-\|\|x-y\|\|^2}$
Inverse Multi-quadratic	$K(x,y) = (\|\|x-y\|\|^2 + c^2)^{-1/2}$
Multi-quadratic	$K(x,y) = (\|\|x-y\|\|^2 + c^2)^{1/2}$
Sigmoid	$K(x,y) = tanh(xy - \theta)$

Table 3 lists training and testing errors for networks found by genetic parameter search, first with Gaussian kernel (which is the most commonly used kernel function), second by kernels optimised by genetic search. In addition, errors obtained by RBF networks are listed. In all cases, Inverse Multi-quadratic kernel function was winning in the evolution. It also in all cases gives better results than Gaussian kernel. Fig. 4 and Fig. 5 shows the evolution of subpopulation sizes. The former is obtained with tournament selection, the latter with roulette-wheel selection. In case of tournament selection, the first 10 generations are enough for Inverse Multi-quadratic kernel function to dominate over population and other kernel functions to die out. In case of roulette-wheel selection, the population stays diverse, the most represented kernel functions are Inverse Multi-quadratic and Gaussian kernel functions.

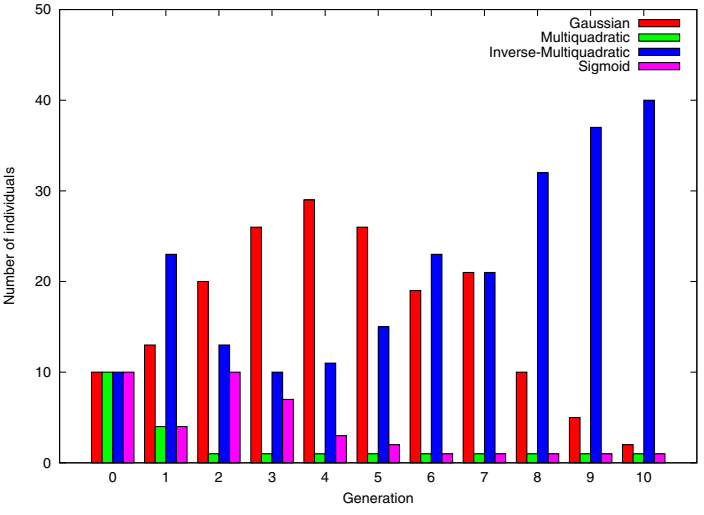


Fig. 4. Numbers of individuals in subpopulations during evolution for the case of tournament selection

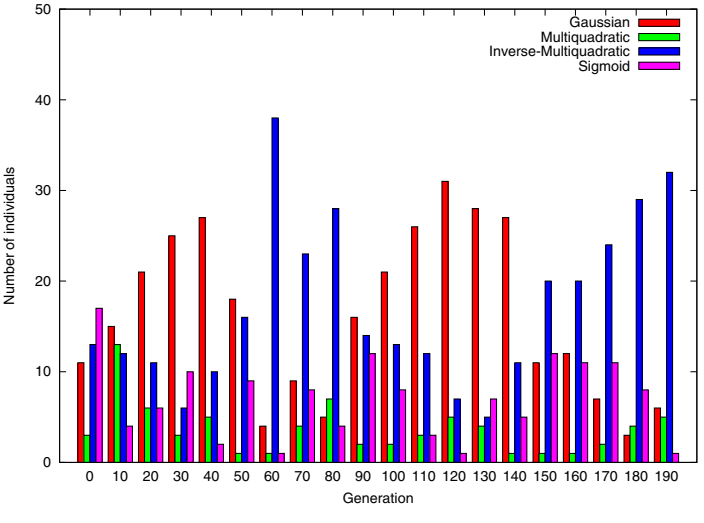


Fig. 5. Numbers of individuals in subpopulations during evolution for the case of roulette-wheel selection

Table 3. Error on training and testing set obtained by RN with Gaussian kernels and RN with metaparameters optimized by genetic parameter search.

	Gaussian kernel		Genetic search			RBF	
Task	E_{train}	E_{test}	E_{train}	E_{test}	Winning Kernel	E_{train}	E_{test}
cancer1	2.29	1.76	1.83	**1.50**	Inverse Multiquadratic	2.31	2.11
cancer2	1.85	3.01	1.41	**2.92**	Inverse Multiquadratic	1.91	3.12
cancer3	2.05	2.78	1.74	**2.54**	Inverse Multiquadratic	1.66	3.19
card1	7.17	10.29	7.56	**10.03**	Inverse Multiquadratic	8.12	10.16
card2	5.72	13.19	6.06	**12.74**	Inverse Multiquadratic	8.05	12.81
card3	5.93	12.68	6.35	**12.28**	Inverse Multiquadratic	6.77	12.09
flare1	0.34	**0.54**	0.35	**0.54**	Inverse Multiquadratic	0.37	0.37
flare2	0.41	**0.27**	0.41	**0.27**	Inverse Multiquadratic	0.41	0.31
flare3	0.39	**0.33**	0.39	**0.33**	Inverse Multiquadratic	0.37	0.38
glass1	3.37	6.99	2.32	**6.13**	Inverse Multiquadratic	5.10	6.76
glass2	3.92	7.74	1.06	**6.79**	Inverse Multiquadratic	4.93	7.96
glass3	4.11	7.36	2.67	**6.27**	Inverse Multiquadratic	5.80	8.06
heartac1	3.39	3.30				2.26	3.69
heartac2	2.08	4.15				1.78	4.98
heartac3	2.52	5.10				1.66	5.81
hearta1	2.58	4.43				3.08	4.36
hearta2	2.12	4.32				3.36	4.05
hearta3	2.59	4.44				3.19	4.29
heartc1	7.73	16.03				6.07	16.17
heartc2	10.67	6.82				7.99	6.49
heartc3	6.88	13.36				7.13	14.35
heart1	8.56	13.70	9.21	**13.55**	Inverse Multiquadratic	9.96	14.05
heart2	7.99	14.15	8.17	**13.88**	Inverse Multiquadratic	6.36	11.67
heart3	5.85	17.03	5.92	**16.85**	Inverse Multiquadratic	6.95	12.02
horse1	2.52	13.31	4.15	**11.77**	Inverse Multiquadratic	10.57	11.96
horse2	1.58	16.12	3.63	**15.22**	Inverse Multiquadratic	10.04	16.80
horse3	2.27	14.62	3.84	**13.53**	Inverse Multiquadratic	9.88	14.56

6 Conclusion

We introduced the method for optimization of metaparameters of RN based on genetic algorithms. The method was demonstrated on experiments and it found better results than RN with Gaussian kernel only.

Though Gaussian kernel function is the most common one, in all cases were the best results obtained by Inverse Multi-quadratic kernel function. This function has similar shape as Gaussian function, which was the second one often presented in the population. Sigmoid and Multi-quadratic kernel functions were not successful.

In our future work we plan to extend genetic search also for composite type of kernel function, combining different kind of kernels together using operator product and sum of kernels [9].

Acknowledgment

This research has been supported by the Grant Agency of the Czech Republic under project no. 201/08/1744.

References

1. Girosi, F., Jones, M., Poggio, T.: Regularization theory and Neural Networks architectures. Neural Computation 2, 219–269 (1995)
2. Kůrková, V.: Learning from data as an inverse problem. In: Antoch, J. (ed.) Computational Statistics, pp. 1377–1384. Physica Verlag, Heidelberg (2004)
3. Poggio, T., Girosi, F.: A theory of networks for approximation and learning. Technical report, Cambridge, MA, USA, A. I. Memo No. 1140, C.B.I.P. Paper No. 31 (1989)
4. Poggio, T., Smale, S.: The mathematics of learning: Dealing with data. Notices of the AMS 50, 536–544 (2003)
5. Tikhonov, A., Arsenin, V.: Solutions of Ill-posed Problems. W.H. Winston, Washington (1977)
6. Mitchell, M.: An Introduction to Genetic Algorithms. MIT Press, Cambridge (1996)
7. Prechelt, L.: PROBEN1 – a set of benchmarks and benchmarking rules for neural network training algorithms. Technical Report 21/94, Universitaet Karlsruhe (September 1994)
8. LAPACK: Linear algebra package, http://www.netlib.org/lapack/
9. Kudová, P., Šámalová, T.: Sum and product kernel regularization networks. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Żurada, J.M. (eds.) ICAISC 2006. LNCS (LNAI), vol. 4029, pp. 56–65. Springer, Heidelberg (2006)

Integrating Personalized and Community Services for Mobile Travel Planning and Management

Chien-Chih Yu

Dept. of MIS, National ChengChi University, Taipei, Taiwan
ccyu@mis.nccu.edu.tw

Abstract. Personalized and community services have been noted as keys to enhance and facilitate e-tourism as well as mobile applications. This paper aims at proposing an integrated service framework for combining personalized and community functions to support mobile travel planning and management. Major mobile tourism related planning and decision support functions specified include personalized profile management, information search and notification, evaluation and recommendation, do-it-yourself planning and design, community and collaboration management, auction and negotiation, transaction and payment, as well as trip tracking and quality control. A system implementation process with an example prototype is also presented for illustrating the feasibility and effectiveness of the proposed system framework, process model, and development methodology.

Keywords: Personalized services, community services, mobile services, travel planning, integrated service framework.

1 Introduction

Personalized services refer to information, communication, transaction, or decision support services that are tailored to meet customers' needs and preferences [1]. Community services, on the other hand, provide functions for customers to form special interest groups, locate people with similar interests, set up community forum and collaborative platform, as well as share information and experiences [6,24]. The rapid advancement of mobile technologies has stimulated the demands for more personalized and community based mobile services in a wide range of application domains such as tourism, healthcare, learning, and government [6,8,10,12,22,24]. In general, mobile services refer to information, communication, and transaction services that are accessed and delivered via mobile communications networks. Customer values identified for mobile services include ubiquitous access, convenience, personalization, localization, productivity enabler, time saving, better tracking, cost reduction, security enhancement, as well as lower prices and special promotions [2,26]. Specifically for the tourism application domain, mobile technologies have been pointed out as one of the most promising innovative technologies for promoting and enhancing the quality of tourism related services, as well as for creating tourists' values [3,6]. Tourists or travelers are typical on-the-move customers that need services with mobility and

B. Papasratorn et al. (Eds.): IAIT 2010, CCIS 114, pp. 202–213, 2010.

ubiquity features. They like to get help in designing, operating, adjusting, and recording personalized travel plans that take into account their needs and preferences in all pre-trip, during the trip and post trip phases [11,29,32]. They also like to seek and share information and experiences with other people of similar tourism interests, get recommendations from community members based on their past travel cases and ratings, and possibly form a tour group and plan their joint trips through social interaction and collaboration [6,9,24].

As a key service in tourism application, tourism and travel planning is a process of searching, selecting, grouping and sequencing destination related products and services including attractions, accommodations, restaurants, and activities [4,5,7,11,17,18,25,27,32]. As more functional support on ubiquitous accessibility, tourism information availability, and community interactivity emerged, the requirements of more personalization, localization, and collaboration services for mobile tourism related planning and transactions increase significantly [4,11,32]. As a result, how to provide and integrate personalized and community services for enhancing and facilitating travel planning and management have become critical issues in tourism related research and practices. However, little progress has been made in this area to improve the performance and management of travel planning processes. Existing commercial mobile services and research works regarding personalized and community based travel planning services are still very limited and primitive. There is still a lack of integrated service framework and design process for guiding the development of personalized and community enabled travel planning services to meet the travelers' needs and preferences. For filling the research gap, this paper aims at proposing an integrated service framework for combining personalized and community functions to support user-centric mobile travel planning and management. Also presented include related system design methods and an example prototype to illustrate the feasibility and effectiveness of the proposed integrated service framework, application process, and system development approach. The rest of this paper is organized as follows. A comprehensive literature review on personalized and community services in mobile tourism applications is provided in section 2. In section 3, the integrated service framework for supporting personalized and community based mobile travel planning and management is presented. Section 4 contains the illustration of a system development process with a prototype. A concluding remark with future research issues are given in the final section.

2 Literature Review

In the following subsections, previous research works related to personalized services as well as community and collaborative services in mobile and tourism application domains, separately and jointly, are reviewed and discussed.

2.1 Personalized Services in Mobile and Tourism Application Domains

In the literature of web-based personalization, major types of personalized services in various application domains include the directory and search services, the selection and recommendation services, the self-planning and customization services, as well as

the auction and negotiation services. In mobile applications, personalized services often take users' location and time as part of the conditions for information search and recommendation [8,12]. As for the tourism application domain, previous research have also circled strong demands of information-intensive and decision-support services that incorporate personalized needs and preferences in all tourism-related searching, decision making, and transaction processes [19,25,29]. For instances, Ricci et al. (2002) present a case-based reasoning (CBR) approach to be adopted in travel recommender systems for assisting users in travel-related information filtering and product bundling [19]. Tomai et al (2005) explore how to use ontologies for assisting tourists in trip planning in a web-based environment by illustrating the building and matching of the users profile and tourism information ontologies [25].

Focusing on issues and methodologies for developing personalized mobile tourism applications, Scherp and Boll (2004) present an approach and example of dynamically generating personalized multimedia content for generic tourist guide based on the tourist's interests and preferences, current location and environment, and mobile device used [25]. Breunig and Baer (2004) collect requirements and present an implementation prototype of a mobile route planning system that is capable of supporting spatial database queries [4]. Adopting multi-agent and semantic web technologies with aims to effectively coordinate and integrate disparate information and service resources anytime and anywhere, Chiu and Leung (2005) propose a ubiquitous tourist assistance system for providing personalized assistances to tourists by taking into account their different preferences and often changing during-he-rip requirements [9]. By presenting location-based applications based on the combination of Global Navigation Satellite System (GNSS) and Geographical Information System, Sadoun and Al-Bayari (2007) show the potential of using positioning techniques for improving location determination, geo-location information navigation, vehicle tracking and route planning services. [21]. For supporting dynamic and context-based trail management, Driver and Clarke (2008) propose an application framework that comprises trail generation and trail reconfiguration point identification modules for effectively managing mobile and context-aware trails [11]. Castillo et al. (2008), to provide supports for planning tourist visits in a city, present an user-oriented adaptive system that is intended to work in portable devices with internet connection and is functionally capable of accessing ontology-based information, capturing and updating a user model about different city visits, selecting a list of interesting places to visit through a case-based reasoning approach, as well as generating a schedule plan by taking into account goals, preferences, places, distances, timetables, and transportation means [7]. To facilitate the design of a personalized route planning system, Niaraki and Kim (2009) present a generic ontology-based architecture using an analytic hierarchical process (AHP) to determine an appropriate impedance model of road segments based on user preferences [18]. Through presenting a prototype developed on the top of Java 2 Micro Edition (J2ME), Kenteris et al. (2009) highlight main advantages and shortcomings in implementing a mobile tourist guide application that enables the automated creation of portable, personalized tourist applications with rich and customized content while minimizes the wireless connectivity requirement of the mobile users [15]. To efficiently and effectively provide various mobile recommendations regarding sightseeing spots, hotels, restaurants, and packaged tour plans, Yu and Chang (2009) propose a system architecture and design methods for facilitating the delivery

of location-based recommendation services to support personalized tour planning based on tourists' current location and time, preferences and needs, as well as constraints and selection criteria [32].

2.2 Community Services in Mobile and Tourism Application Domains

In research addressing community and collaborative services, major service types mentioned include profile management, resource management, community management, conflict management, search and notification, activity and event arrangement, group communications and messaging, as well as meeting and conferencing [6,10,16,24,28,30,31]. For research involving the mobile application environment, features and issues such as mobility, ubiquity, and localization are emphasized and addressed for conveniently activating community and collaborative services independent of location, time, and connectivity constraints [10,16,24]. As examples, in a three layered service architecture for supporting mobile teamwork, Kirda et al. (2002) organize access control, user management, community management, artifact management, repository, messaging, subscription, and distributed search as components of the teamwork services [16]. Addressing the need of mobile share workspaces for ubiquitous collaboration, Divitini et al. (2004) present an experimental collaboration service platform in which services offered enable users to create and maintain profiles of users, activities, and resources, to dynamically configure these entities, and to maintain the presence model [10]. Describing the community platform as an innovative value added service system for supporting mobile coordination of individuals, Schubert and Hampe (2006) present a group of value services to community members of the leisure industry that include searching personalized information specials, seeking leisure partners, and coordinative arrangement of leisure events [24].

As for research in the community-enabled electronic and mobile tourism domain, only initial studies have been delivered [6,28,30]. Wang et al. (2002) define virtual tourist community and discuss its implications for tourism marketing [28]. By conducting an empirical study on members of online travel communities, Wu and Chang (2005) report findings about how interactivity and trust affect the experience of flow and how flow affects transaction intensions of these members [30]. While targeting on mobile applications, Carlsson et al. (2008) present a mobile community based service that incorporates three key elements, namely online tourist communities, mobile blogging, and mobile social networks to facilitate information search and experience sharing through social interactions among tourists co-located at a given destination [6].

From the above literature review, it can be seen that although personalized services, community services, mobile services, and tourism services have been noted as strongly related fields for research and practices, very few have provided integrative views and frameworks that incorporate the featuring concepts and methodologies to facilitate hybrid personalized and community services for mobile tourism applications, especially for mobile travel planning. Among these rare works, Yu (2005) propose a functional framework and design process for building customer-oriented decision support systems to jointly deliver personalized and community tourism services [31]. Major decision support functions include personalized data and model management, information search and navigation, product/vendor evaluation

and recommendation, do-it-yourself travel planning and design, community and collaboration management, auction and negotiation, as well as trip tracking and quality control. However, only a limited scope of mobility features has been discussed. On the other hand, focusing on designing location based tourism services, Kansa and Wilde (2008) point out that the central requirements for empowering tourists to filter and augment their travel experiences, and to co-create and engage generated content and experiences with peers include ubiquitous recommendation, augmentation of travel realities, as well as peer production of tourist realities [14]. Nevertheless, decision functions regarding community and collaboration support for travel planning have not been fully explored. In summary, previous research regarding personalized services and/or collaborative services in mobile tourism applications usually dealt with problems of limited scopes and focused only on specific functions and/or techniques. The needs of an integrated architecture and design process to incorporate personalized and community services for supporting mobile individual as well as coordinated group travel planning are significant.

3 The Integrated Service Framework and Process Model

Through an ideal integrated personalized and community-based mobile travel planning support system, users should be able to create and maintain personal profiles, to search location and context aware information, to view geo-information and maps with GIS support, to share information, experiences, and resources with community members, to receive information and recommendations about point-of-interests and tour plans based on their needs, preferences, and conditions, as well as on experiences and ratings of other users with similar conditions and interests, to plan a personalized trip schedule, to seek travel partners and to collaboratively plan a group travel plan, to search and select tourism operators based on auction and negotiation, to track during-the-trip positions and conditions for assuring safety and quality of the travel plan, to rate point-of-interests and travel plans, as well as to store post-trip travel plans as cases for future references. To develop such a powerful personalized and community-based mobile tourism support system for travel planning and management, major information management and decision support functions to be specified in the application level include mobile personalized profile management, mobile information search and notification, mobile evaluation and recommendation, mobile do-it-yourself planning and design, mobile community and collaboration management, mobile auction and negotiation, mobile transaction and payment, as well as mobile trip tracking and quality control. Both the specific point-of-interests and travel plan that bundles attractions, activities, durations and distances with sequential orders can be acquired using the location-based and context-aware personalized recommendation services with map-based positioning and visualization facilities. Functional requirements of the back end system level include data base, model base, and knowledge base management, as well as ontology and case base management. Descriptions of the integrated service framework and process model for guiding the development and operation of the ideal system are provided below.

3.1 The Integrated Service Framework

The integrated service framework of personalized and community-based mobile travel planning and management support system is shown in Figure 1. Specifications of associated service functions are as follows.

Mobile Personalized Profile Management: Functions in this group enable users to create and maintain their personal profiles including basic personal information and personalized travel preferences, previous visits to cities and places of interest such as sightseeing spots, restaurants, and hotels with satisfactory ratings, as well as past implemented travel plans with ratings. Users can also specify their personalized evaluation criteria for selecting tourism-related products, services, agents, operators, or tour plans. Also included are facilities for users to sign up to specific community groups, to create and manage personalized link lists of frequently accessed tourism sites or favorite points of interest, as well as to record and maintain their travel histories by using blogs, photo albums, and other textual/multimedia documents.

Mobile Information Search and Notification: These functions allow users to search, retrieve, navigate and browse requested tourism information such as destinations and events, weather and traffic conditions, airline or train schedules, nearby restaurants or shopping centers, hotels with available rooms, travel agencies and group packaged tours, as well as travel experiences and resources of other community users, etc. Users can also receive notification and relevant information such as airplane boarding time or hotel check in/check out time, promotion or other special events in nearby shopping malls, etc in a pushed manner.

Mobile Evaluation and Recommendation: These grouped functions support users for specifying their needs, preferences and criteria, as well as current location, time and constraints, and then for activating the evaluation process to generate system recommended tourism products, services, points of interest, or package tours that closely match users' needs and preferences. Users' needs, preferences and evaluation criteria can be specified by accessing data from the user profiles or by directly inputting data as specific instances. Intelligent decision support functions with rule sets and decision models are used for recommending travel agencies' matched package tours, while case-based reasoning is used to recommend community members' similar travel cases.

Mobile Do-It-Yourself Planning and Design: Main functions of this service group provide a simple interface to allow users interactively specifying cities, attractions, restaurants and hotels in a sequential date basis to design and form personalized travel plans. Users can start the DIY process from scratch by picking, bundling, and sequencing chosen cities, sightseeing spots, restaurants and hotels in a daily basis, or from a retrieved case of implemented travel plan with similar features.

Mobile Community and Collaboration Management: Using the basic community services, users are capable of locating members of similar travel interests, forming special tourism interest groups, setting up tourism-related community forums and communication channels, as well as sharing travel experiences and resources with peers. Collaborative service functions allow users to propose their personal travel plans to the community, to exchange ideas and collaboratively design alternative travel plans

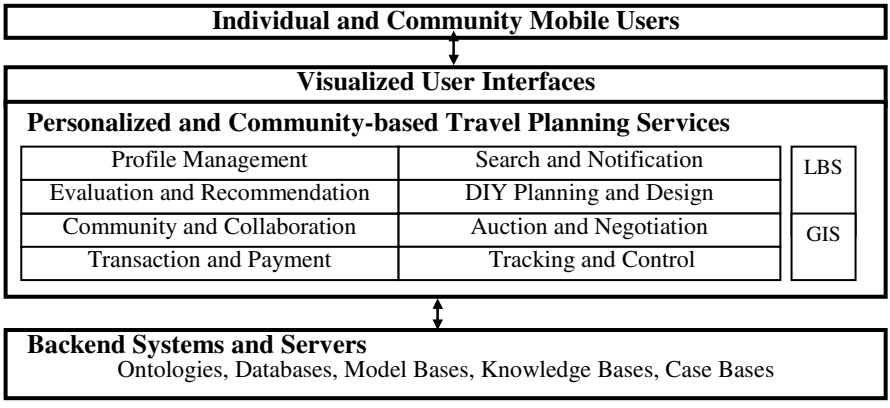


Fig. 1. The integrated service framework for travel planning and management

with interested members, to resolve conflicts, vote for proposals, and select ones with the highest votes as the commonly accepted group travel plans for implementation. The initial proposal can be a result of the DIY planning process or a previously implemented case recommended by the system based on the CBR technique.

Mobile Auction and Negotiation: This functional group provides a dynamic and competitive pricing environment for users to issue tourism requests with specified needs and preferences, and to launch reverse auction sessions that call for tourism service providers to bid on the posted individual or group travel plans. The submitted bids are evaluated and those with top satisfactory cost/benefit levels are selected as candidates for contract negotiation. The negotiation function allows users to negotiate with chosen tourism service providers by exchanging modified terms and contracts.

Mobile Transaction and Payment: This functional group allows users to use mobile devices for actually booking and issuing payments to selected tour plans, for hotel or restaurant reservation, or for purchasing tourism related products such as train tickets, museum directories, or destination-specific souvenirs.

Mobile Trip Tracking and Quality Control: This group of functions provides mechanisms for travelers, their families, and the travel agents to track during-the-trip locations and situations, and if necessary, make changes to the operating travel plans by rearranging routes, points of visit, restaurants, or hotels for assuring the safety and quality levels of the travel plans.

Associated Location-Based Services and Map Services: These services take into account traveler's current location and time for finding nearby point-of-interests or planning an area trip. The spots to be visited as well as their routes are shown using Google Map.

3.2 The Integrated Process Model

Based on the proposed integrated service framework, the process model for mobile travel planning and management is depicted in Figure 2 with associated steps being described below.

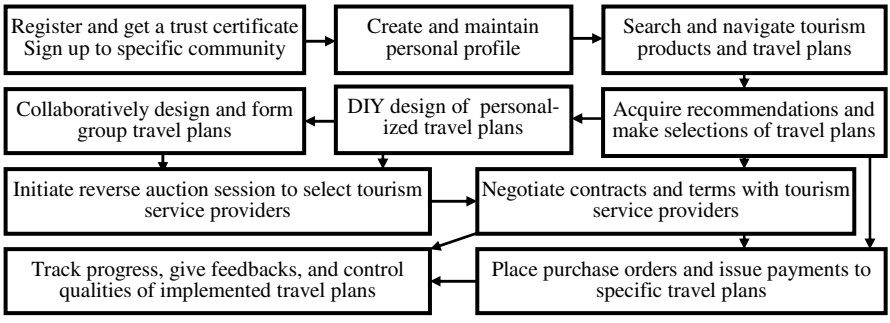


Fig. 2. The integrated process model for travel planning and management

1. Users can register as a member and get trust certificate, and also sign up to join specific communities.
2. Users create and maintain personal profiles that contain categorized data about users' basic information, personalized travel related needs, preferences, and constraints, past travel histories with implemented travel plans, photo albums, and text descriptions, link lists of favorite web pages, as well as personalized evaluation criteria for selecting tourism products and services.
3. Users search and navigate tourism-related information such as destinations and accommodations, attractions and features, package tours and travel agencies, etc., as well as to access other users' experiences, ratings and resources related to specific destinations and accommodations. Users also receive product or event notifications relevant to their needs, preferences, and conditions.
4. Users activate the evaluation and recommendation procedure using specified criteria to choose tourism products and vendors, or past travel cases that meet their needs and preferences with specified matching or similarity levels.
5. Users design personalized travel plans when there is no existing package tours or travel cases with the pre-specified satisfaction levels recommended by the system.
6. Users can join or form communities of special interests, locate and organize users of similar interests to exchange experiences and ideas, and design collaboratively with community members to develop commonly accepted group tour plans.
7. Users propose requests and initiate reverse-auction sessions for inviting tourism vendors and services providers to bid on proposed individual and/or group tour plans, as well as to evaluate and select those with best bids as candidates
8. Users negotiate terms and contracts of the selected tour plans with chosen tourism service providers such as travel agencies and tour operators.
9. Users can book and pay for selected tour plans, and receive new and just-in-time information in pull and push manners.
10. Users can track the progress and control qualities during the tour operations of the contracted travel plans.

With planning and decision support services to this extent, all pre-, during, and post phases of the tourism and travel related decision-making process can be fully supported through successfully incorporated personalized and community services.

4 The Design Method and Prototype System

For developing the integrated mobile travel planning and management support system, the scopes of system design include architecture design, process design, user interface design, presentation design, database design, model base design, knowledge base design, as well as the ontology and case base design, etc. In the following, design considerations for ontologies, databases, model bases, knowledge bases, and case bases, as well as an implemented prototype are provided.

4.1 Ontology, Database, Model Base, Knowledge Base, and Case Base Design

For constructing tourism ontology and user ontology, as well as for designing the system database, model base, knowledge base, and case base, an unified Object-Oriented (OO) model design approach is adopted to create the integrated conceptual data model. The OO model is then translated into internal models for implementing associated databases, model bases, knowledge bases, and case bases that conform with the ontologies. For instance, an entity-relationship (ER) model is derived and further transformed into a relational model for physical database implementation. Figure 3 presents partially the OO data model in which identified objects include Traveler Profile, Tourism Information, Travel Plan, and Recommendation Process. The Traveler Profile object composes of Needs and Preferences (N&P), Search Constraint and Criteria, as well as Current Location and Time objects. The Needs and Preferences object is further classified into Sightseeing N&P, Restaurant N&P, and Hotel N&P objects. The Tourism Information object aggregates three sub-class objects including Sightseeing Spots Information, Restaurant Information, and Hotel Information objects, as well as packaged travel plan and past case objects. The Recommendation Process object has several component Match Process objects including Ontology Match Process, Case-Based Reasoning Process, and Knowledge and Model Computing Process objects. Each Match Process object contains specific Process Input, Process Execution, and Process Output objects. For instance, the Process Execution objects for the Knowledge and Model Computing Process comprise Model Input, Recommendation Model, Model Output, Rule Input, Recommendation Rule, and Rule Output objects. The recommended Travel Plan object consists of Sightseeing Spot Selection, Restaurant Selection, and Hotel Selection objects as the components.

4.2 The Prototype System

A prototype system that provides the integrated personalized and community services for mobile travel planning is developed in an environment using system and application software such as Windows XP, Microsoft IIS Web Server 5.1, .NET Framework 2.0, and Microsoft SQL Server 2005, as well as ASP.NET 2.0, Web Services, and Google Map API 2.0. The CHT Windows Mobile 5.0 Smart Phone Emulator is used as the client-side emulator. The prototype system allows travelers to access desired personalized and community services for travel planning via PDAs or smartphones. Sightseeing spots, restaurants, and hotels can be selected, or a packaged

travel plan (for individual or group) can be generated based on travelers' needs, preferences, constraints, conditions, and specified evaluation criteria. Figure 4 shows example mobile phone screen shots of the prototype system including updating the traveler profile, needs setting for search travel plans, presenting the recommended travel plan, and showing a map with the route and sequential visits.

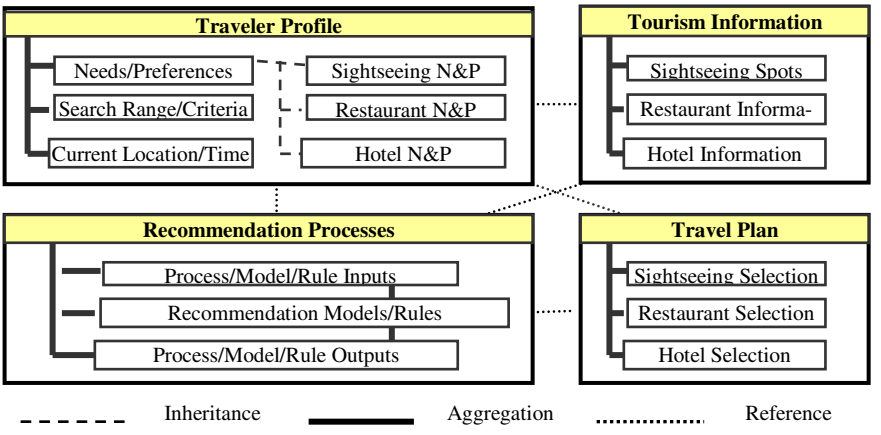


Fig. 3. The Object-Oriented conceptual data model

Fig. 4. Output screenshots of the prototype system

The prototype system has been evaluated by 32 graduate and EMBA students who have both travel planning and mobile application usage experiences. Based on a 5-point Likert scale (1 as strongly disagree and 5 as strongly agree), the average scores of the 4 performance measures including usefulness, ease of use, satisfaction, and intention for future use are 4.09, 4.14, 3.79, and 3.86 respectively with the overall average score being above 3.97. These outcomes validate the feasibility and effectiveness of the proposed framework, process, and design methods.

5 Concluding Remark and Future Works

In this paper, we present an integrated service framework, process model, and design methods for guiding the development of personalized and community-support mobile travel planning and management services. A system prototype is also developed and tested to validate the feasibility and effectiveness of the proposed approach with satisfaction. Future research works include evaluating efficiency and effectiveness of the system using real world cases, comparing the performance of ontology matching, case-based reasoning, with model/knowledge computing mechanisms using various scenarios, as well as integrating web 2.0 technology into the system.

References

1. Adomavicius, G., Tuzhilin, A.: Personalization Technologies: A Process-Oriented Perspective. Communications of the ACM 48(10), 83–90 (2005)
2. Anckar, B., D'Incau, D.: Value Creation in Mobile Commerce: Findings from a Consumer Survey. J. of Info. Technology Theory and Application 4(1), 43–64 (2002)
3. Buhalis, D., Law, R.: Progress in Information Technology and Tourism Management: 20 Years on and 10 Years after the Internet—The State of eTourism Research. Tourism Management 29(4), 609–623 (2008)
4. Breunig, M., Baer, W.: Database Support for Mobile Route Planning Systems. Computers, Environment and Urban Systems 28(6), 595–610 (2004)
5. Cardoso, J.: Developing Dynamic Packaging Systems Using Semantic Web Technologies. Transactions on Information Science and Applications 3(4), 729–736 (2006)
6. Carlsson, C., Walden, P., Yang, F.: Travel MoCo - A Mobile Community Service for Tourism. In: Proceedings of the 7th International Conference on Mobile Business, pp. 49–58 (2008)
7. Castillo, L., et al.: SAMAP: An User-Oriented Adaptive System for Planning Tourist Visits. Expert Systems with Applications 34(2), 1318–1332 (2008)
8. Chen, M., Zhang, D., Zhou, L.: Providing Web Services to Mobile Users: The Architecture Design of an M-Service Portal. International Journal of Mobile Communications 3(1), 1–18 (2005)
9. Chiu, D.K.W., Leung, H.F.: Towards Ubiquitous Tourist Service Coordination and Integration: a Multi-Agent and Semantic Web Approach. In: Proceedings of the 2005 International Conference on Electronic Commerce, pp. 574–581 (2005)
10. Divitini, M., Farshchain, B.A., Samset, H.: UbiCollab: Collaboration Support for Mobile Users. In: Proceedings of the 2004 ACM Symposium on Applied Computing, pp. 1191–1195 (2004)
11. Driver, C., Clarke, S.: An Application Framework for Mobile, Context-Aware Trails. Pervasive and Mobile Computing 4(5), 719–736 (2008)
12. Guo, X., Lu, J.: Intelligent E-Government Services with Personalized Recommendation Techniques. International Journal of Intelligent Systems 22, 401–417 (2007)
13. Kakaletris, G., et al.: Designing and Implementing an Open Infrastructure for Location-Based, Tourism-Related Content Delivery. Wireless Personal Communications 30(2-4), 153–165 (2004)
14. Kansa, E.C., Wilde, E.: Tourism, Peer Production, and Location-Based Service Design. In: Proceedings of the 2008 IEEE International Conference on Service Computing, pp. 629–636 (2008)

15. Kenteris, M., Gavalas, D., Economou, D.: An Innovative Electronic Tourist Guide Application. Pervasive and Ubiquitous Computing 13(2), 103–118 (2009)
16. Kirda, E., Fenkam, P., Reif, G., Gall, H.: A Service Architecture for Mobile Teamwork. In: Proceedings of the 14th International Conference on Software Engineering and Knowledge Engineering, pp. 513–518 (2002)
17. Knoblock, C.: Agents for Gathering, Integrating, and Monitoring Information for Travel Planning. IEEE Intelligent Systems 17(6), 63–64 (2002)
18. Niaraki, A.B., Kim, K.: Ontology Based Personalized Route Planning System Using a Multi-Criteria Decision Making Approach. Expert Systems with Applications 36(2p1), 2250–2259 (2009)
19. Ricci, F., Arslan, B., Mirzadeh, N., Venturini, A.: ITR: A Case-based Travel Advisory System. In: Craw, S., Preece, A.D. (eds.) ECCBR 2002. LNCS (LNAI), vol. 2416, pp. 613–627. Springer, Heidelberg (2002)
20. Ricci, F., Nguyen, Q.N.: Acquiring and Revising Preferences in a Critique-Based Mobile Recommender System. IEEE Intelligent Systems 22(3), 22–29 (2007)
21. Sadoun, B., Al-Bayari, O.: Location Based Services Using Geographical Information Systems. Computer Communications 30(16), 3154–3160 (2007)
22. Scherp, A., Boll, S.: Generic Support for Personalized Mobile Multimedia Tourist Applications. In: Proceedings of the 2004 ACM International Conference on Multimedia, pp. 178–179 (2004)
23. Schilke, S.W., Bleimann, U., Furnell, S.M., Phippen, A.D.: Multi-Dimensional-Personalisation for Location and Interest-Based Recommendation. Internet Research 14(5), 379–385 (2004)
24. Schubert, P., Hampe, J.F.: Mobile Communities: How Viable are Their Business Models? An Exemplary Investigation of the Leisure Industry. Electronic Commerce Research 6(1), 103–121 (2006)
25. Tomai, E., Spanaki, M., Prastacos, P., Kavouras, M.: Ontology Assisted Decision Making – A Case Study in Trip Planning for Tourism. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM-WS 2005. LNCS, vol. 3762, pp. 1137–1146. Springer, Heidelberg (2005)
26. Varshney, U.: Business Models for Mobile Commerce Services: Requirements, Design, and the Future. IT Professional 10(6), 48–55 (2008)
27. Vila, M., Costa, G., Rovira, X.: The Creation and Use of Scorecards in Tourism Planning: A Spanish Example. Tourism Management 31(7), 232–239 (2010)
28. Wang, Y., Yu, Q., Fesenmaier, D.R.: Defining the Virtual Tourist Community: Implications for Tourism Marketing. Tourism Management 23(4), 407–417 (2002)
29. Werthner, H., Ricci, F.: E-commerce and Tourism. Communication of the ACM 47(12), 101–105 (2004)
30. Wu, J.J., Chang, Y.S.: Towards Understanding Members' Interactivity, Trust, and Flow in Online Travel Community. Industrial Management and Data Systems 105(7), 937–954 (2005)
31. Yu, C.C.: Personalized and Community Decision Support in eTourism Intermediaries. In: Andersen, K.V., Debenham, J., Wagner, R. (eds.) DEXA 2005. LNCS, vol. 3588, pp. 900–909. Springer, Heidelberg (2005)
32. Yu, C.C., Chang, H.P.: Personalized Location-based Recommendation Services for Tour Planning in Mobile Tourism Applications. In: Di Noia, T., Buccafurri, F. (eds.) E-Commerce and Web Technologies. LNCS, vol. 5692, pp. 38–49. Springer, Heidelberg (2009)

Author Index