Real-Time Image Recognition and Path Tracking of a Wheeled Mobile Robot for Taking an Elevator

Jih-Gau Juang[1], Chia-Lung Yu[2], Chih-Min Lin[3], Rong-Guan Yeh[4], Imre J. Rudas[5]

[1,2]Department of Communications, Navigation and Control Engineering, National Taiwan Ocean University, Keelung, 202, Taiwan, e-mail: jgjuang@ntou.edu.tw

[3,4]Department of Electrical Engineering, Yuan Ze University, Chung-Li, Taoyuan, 320, Taiwan, e-mail: cml@saturn.yzu.edu.tw

[5]Óbuda University, Bécsi út 96/B, H-1034 Budapest, Hungary, e-mail: rudas@uni-obuda.hu

Abstract: This paper aims to design a wheeled mobile robot for path tracking and for automatically taking an elevator by integrating multiple technologies, including image processing using hue-saturation-value color space, pattern recognition using the adaptive resonance theory (ART), and robot control using a fuzzy cerebellar model articulation controller (FCMAC). The ART is used to recognize the button of the elevator. It is an effective competitive learning rule for figure identification. The FCMAC is used for path tracking. Experimental results demonstrate the proposed control system can drive the wheeled robot to take the elevator automatically.

Keywords: Pattern recognition; Adaptive resonance theory; Fuzzy CMAC; Wheeled mobile robot

1 Introduction

With the progress of human civilization, as well as lifestyle changes, intelligent robots are gaining influence in human daily life, such as providing entertainment, life safety, health and other aspects of the service. Intelligent robots integrate mechanics, electronics, automation, control, and communications technologies. Various types of robots have been developed in recent years. For a variety of needs, the development of a robot system combines many advanced theories, such as path planning, visual image processing technology, body positioning, obstacle avoidance techniques, and arm control. In recent years, the wheeled mobile robot

(WMR) has been frequently discussed in mobile robot researches. It has some advantages, such as easy control, high-speed mobility, and energy storage capacity [1, 2], which are better than for the legged robot. WMR has been successfully applied to path planning, parking control, obstacle avoidance, navigation, object tracking, cooperation of multiple robots, etc.

Many researchers have tried to convert expertise into the robotic system so that the human-robot interaction in daily life can be more harmonious. Zhao and Be-Ment [3] addressed six common kinds of three-wheeled robot models and their ability of control. Leow et al. have analyzed the kinetic model of an all-direction wheeled robot [4]. Zhong used an omni-directional mobile robot on a map building application [5]. Chung *et al.* [6] utilized two wheels and different speeds for position control. Shi *et al.* installed multiple sensors on a WMR and utilized fuzzy reasoning value to control the WMR [7]. Path planning and fuzzy logic controller have been developed to make a WMR track a desired path [8]. In the development of the intelligent versatile automobile, the automobile used ultrasound sensors to help it search the correct reflective position in an unknown environment [9, 10].

The cerebellar model articulation controller (CMAC) can be thought of as a learning mechanism to imitate the human brain [11]. By applying the fuzzy membership function into a CMAC, a fuzzy CMAC has been created [12, 13]. The advantages of a fuzzy CMAC over neural networks (NNs) have been presented [14, 15]. Thus, a fuzzy CMAC will be designed for the control scheme of a WMR.

This study aims to design an intelligent control strategy for a wheeled robot to track an assigned path and to take an elevator automatically so that it can be applied for the service of document delivery in a company. This study applies a fuzzy CMAC to a wheeled mobile robot to track a path and uses adaptive resonance theory [16] to identify an elevator button. Since the WMR is a nonlinear model, a dynamic equation is needed for system analysis and control regulation design. The fuzzy controller is one of the most prevalent controllers for nonlinear systems because the control scheme of a fuzzy controller is simple and flexible. A localization system called StarGazer [17] is used to provide coordinate locations for the WMR. For the fuzzy controller, its inputs are the coordinate and the heading angle. Therefore, the WMR can track a preset path accurately. A neural network system based on the adaptive resonance theory (ART) is utilized for pattern recognition. The neural network can identify the elevator button quickly and correctly. The contributions of this study are: (1) that an intelligent control scheme is proposed to control a WMR to take elevator automatically; (2) that the ART is utilized to recognize elevator buttons successfully; (3) that path planning is obtained by a localization system; and (4) that trajectory control is achieved by a fuzzy CMAC.

2 System Description

The experiments are performed on the Ihomer WMR [18], as shown in Fig. 1.

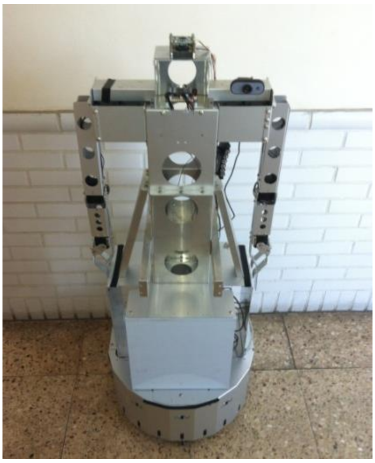


Figure 1
Ihomer wheeled mobile robot (WMR)

A StarGazer localization system is installed on top of the WMR to transmit and receive the infrared signals that are used to obtain the coordinate location and heading angle of the WMR. There are six ultrasonic sensors, of make Devantech SRF05, and the detecting range is from 1 cm to 4 m. The SRF05 sonar sensor module is easy to connect with an ordinary control panel, such as BASICX, etc. The sonar frequency is 40 KHz, the current is 4 mA, and the voltage is 5 V. Assume the condition that the WMR is located on the Cartesian coordinate system (global coordinate system) as shown in Fig. 2, which has no lateral or sliding movement. The WMR represents a vector with nonlinear terms on the two dimensional Cartesian coordinate system carrying both head direction and lateral direction information. The WMR belongs to a nonholonomic system [19]. The movement change of WMR on a global coordinate is

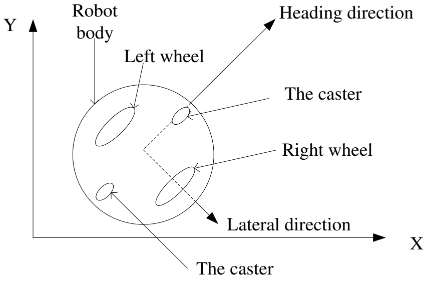


Figure 2
WMR coordinate diagram

$$\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \end{pmatrix} = g_1(q)v + g_2(q)w = \begin{pmatrix} \cos\theta \\ \sin\theta \\ 0 \end{pmatrix} v + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \omega \tag{1}$$

If $\Delta t << 1$, (1) can be replaced by (2)

$$\begin{bmatrix} x_k(i+1) \\ y_k(i+1) \\ \theta(i+1) \end{bmatrix} = \begin{bmatrix} x_k(i) \\ y_k(i) \\ \theta(i) \end{bmatrix} + \begin{bmatrix} \cos\theta & 0 \\ \sin\theta & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v \\ \omega \end{bmatrix} \Delta\mathrm{t} \tag{2}$$

Definitions of all variables are given as follows:

v: Speed of the center of the body

w : Angular speed of the center of the body

r : Radius of the wheel

d : Distance between the left wheel and the right wheel

$x_k(i+1)$: Next x-axis coordinate of the center of the body on global coordinate system

$y_k(i+1)$: Next y-axis coordinate of the center of the body on global coordinate system

$x_k(i)$: X-axis coordinate of the center of the body on global coordinate system

$y_k(i)$: Y-axis coordinate of the center of the body on global coordinate system

Δt : Sample time of every movement of the WMR

w_l : Left-wheel speed of the WMR

w_r : Right-wheel speed of the WMR

In order to obtain the relative distance, we change the target position and the WMR position from the global coordinate system to the local coordinate as shown in Fig. 3, which is between the center of the WMR and the expected position.

$$\begin{bmatrix} x_t' \\ y_t' \end{bmatrix} = \begin{bmatrix} \sin\theta & -\cos\theta \\ \cos\theta & \sin\theta \end{bmatrix} \begin{bmatrix} x_t - x_k \\ y_t - y_k \end{bmatrix} \tag{3}$$

The WMR center relative distance and expected position (error distance) is

$$d_e = \sqrt{x_t'^2 + y_t'^2} \tag{4}$$

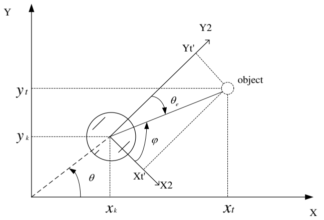


Figure 3

Coordinate of relative position of the WMR and expected position

The relative angle is defined as the included angle between the heading direction of the WMR and the expected position. The local coordinates of the WMR obtained by (5) through the trigonometric function can be calculated.

$$\varphi = \tan^{-1}\left(\frac{y_t'}{x_t'}\right) \tag{5}$$

The error angle of the heading direction and expected position is calculated by (6).

$$\theta_e = 90 - \varphi \tag{6}$$

Definitions of all variables are given as follows:

X: X-axis of the global coordinate system

Y: Y-axis of the global coordinate system

X2: X-axis of the body coordinate system

Y2: Y-axis of the body coordinate system

Xt: The x-axis coordinate of the expected position on the global coordinate system

Yt: The y-axis coordinate of the expected position on the global coordinate system

*X*k: The x-axis coordinate of the center of the body on the global coordinate system

*Y*k: The y-axis coordinate of the center of the body on the global coordinate system

Xt': The x-axis coordinate of the expected position on the body coordinate system

Yt': The y-axis coordinate of the expected position on the body coordinate system

θ: The included angle between the center of the body and the x-axis of the global coordinate system

θe: The relative angle of the heading direction of the WMR and the expected position (error angle)

φ: The included angle between the lateral direction of the WMR and the expected position

de: The relative distance between the WMR and the expected position (error distance)

3 Image Process

The HSV color space is used. The color is made up of its hue, saturation, and brightness. Hue describes where the color is found in the color spectrum and the shade of the color. Red, yellow and purple are used to describe the hue. The saturation describes how pure the hue is with respect to a white reference. The HSV color of external light can improve the identifiable rate of color and the speed of image process [20].

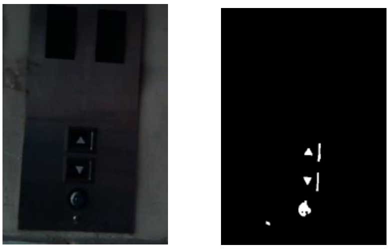
The RGB image of the elevator buttons is shown in Fig. 4(a). The RGB color space values are transferred to the HSV color space values [21] with the following equations:

$$H = \cos^{-1}\frac{(R-G)+(R-B)}{\sqrt{(R-B)^2+(R-B)(G-B)}}, B \le G$$

$$H = 2\pi - \cos^{-1}\frac{(R-G)+(R-B)}{\sqrt{(R-B)^2+(R-B)(G-B)}}, B > G \tag{7}$$

$$S = \frac{\max(R,G,B)-\min(R,G,B)}{\max(R+G+B)}, \quad V = \frac{\max(R,G,B)}{255} \tag{8}$$

V plane is used first. Because the figure that needs identifying remains largely unchanged, a threshold is set to the image and then this image can be processed to become a binary case. Then, open and closed image filtering is used to make the image clearer. This is a very common way of filtering. The processed image is shown in Fig. 4(b).



(a) (b)

Figure 4

The elevator buttons and the processed images

Next, the target image will be identified by the ART method. The ART neural network is shown in Fig. 5.

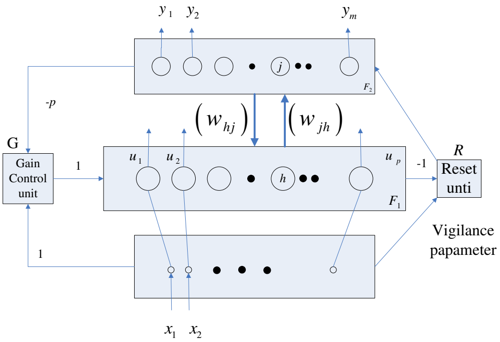


Figure 5
ART neural network

Procedures of the ART neural network are shown as follows:

Step 1: Set all neurons with the initial bounded values.

Step 2: Present the input pattern to the network. Note that the first image pattern is set to the first neuron as the winner and is assigned to class 1, and then skip to Step 6.

Step 3: Enable the neurons that have won and have been assigned to certain classes.

Step 4: According to the following criteria, to find the closest class with the input vector, the so-called ”closest” refers to the “similarity”. The similarity measurement is defined as (the first rating standard):

$$S_1(\underline{w}_j,\underline{x}) = \frac{\underline{w}_j \cdot \underline{x}}{\beta + \left\|\underline{w}_j\right\|_1} = \frac{\underline{w}_j^T \underline{x}}{\beta + \sum_{i=1}^{p} w_{ij}} \tag{9}$$

Step 5: From Step 4, select the winner that has the largest S_1. Assume the j^{th} neuron is the winner, then the second similarity standard is applied to measure the victorious neuron samples stored in the neurons. The second similarity measurement is defined as the evaluation standard (the second rating standard):

$$S_2(\underline{w}_j,\underline{x}) = \frac{\underline{w}_j \cdot \underline{x}}{\left\|\underline{x}\right\|_1} = \frac{\underline{w}_j^T \underline{x}}{\sum_{i=1}^{p} x_i} \tag{10}$$

When $S_2(\underline{w}_j,\underline{x}) \geq \rho$ (ρ is appositive constant and is less than 1, and is called vigilance parameter), the w_j and x can be regard as very similar, then we can perform Step 6; otherwise the j^{th} neuron is disabled, go back to Step 4, find the next winning neuron.

Step 6: Adjust the weights of the winner neuron *j* by following equation:

$$\underline{w}_j(n+1) = \underline{w}_j(n) \ \text{AND} \ \underline{x}\sqrt{b^2 - 4ac} \tag{11}$$

The output of neuron *j* represents the input pattern $\underline{x}$ is recognized as class *j*. Go back to Step 2 to re-accept next new input pattern.

4 Control Scheme

In this study, the control scheme is divided into two parts. The first part is path tracking and the other part is a process to control the robot to push and check the elevator button. A fuzzy CMAC is developed to realize the path tracking control. A linear controller is used to control the WMR in the push-and-check button mission, and another linear controller is used to control the arm's position by the target point with image processing. The control sequence is shown in Fig. 6.

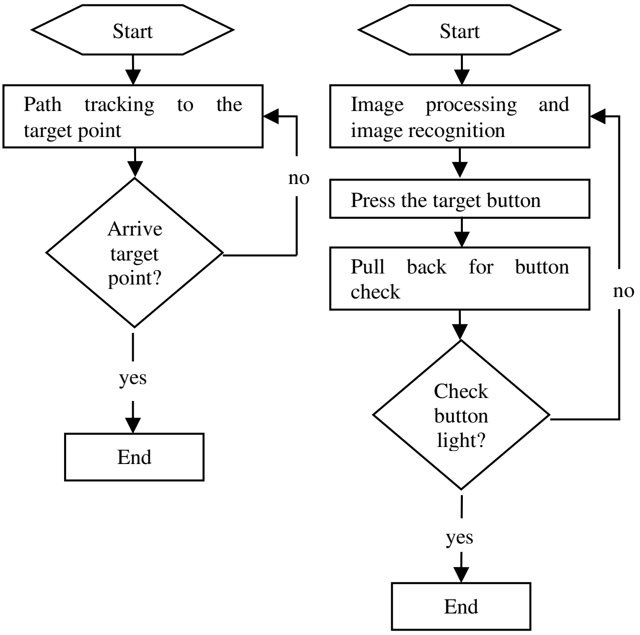

Start
Path tracking to the target point
Arrive target point?
no
yes
End
Start
Image processing and image recognition
Press the target button
Pull back for button check
Check button light?
no
yes
End

Figure 6
Flowchart of the control sequence

A Path Planning

In this study, a localization system is applied in path planning. In order to make the path clear, we need to paste landmarks on the ceiling every 150 centimeters and even in the elevator. With the coordinates obtained by the localization system, we can plan the desired path for the WMR, as shown in Fig. 7.

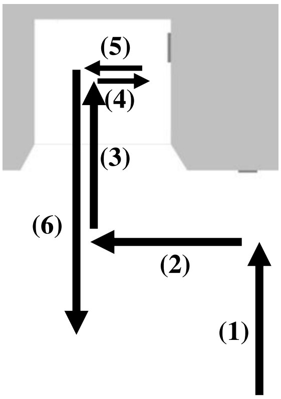


Figure 7
The planning paths

B Fuzzy CMAC Control

In order to take the elevator automatically, the WMR uses the fuzzy CMAC to track the desired path and uses image processing and image recognition to find the elevator button. The considered FCMAC is shown in Fig. 8. Compared with a neural network, the FCMAC has a faster learning rate and better local generalization ability. The processing algorithms of the FCMAC are described as follows:

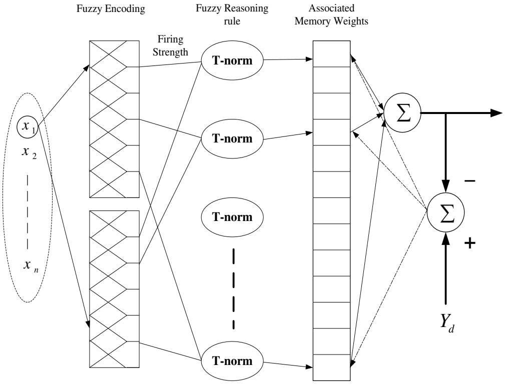


Figure 8
Conventional FCMAC structure

Step 1:

Quantization (*X*→*S*): *X* is an n-dimension input space. For the given, $x = [\, x_1 \;\; x_2 \,, \dots , x_n \,]^T$, $s = [\, s_1 \;\; s_2 \,, \dots , \; s_n \,]^T$ represents the quantization vector of *x*. It is specified as the corresponding state of each input variable before the fuzzification.

Step 2:

Associative Mapping Segment (*S*→*C*): It is to fuzzify the quantization vector which is quantized from *x*. After the input vector is fuzzified, the input state values are transformed to "firing strength" based on the corresponding membership functions.

Step 3:

Memory Weight Mapping (*C*→*W*): The i^{th} rule's firing strength in the FCMAC could be computed as:

$$C_j(x) = c_{j1}(x_1) * c_{j2}(x_2) * .. c_{jn}(x_n) = \prod_{i=1}^{n} c_{j1}(x_i) \tag{12}$$

where $c_{j_i}(x_i)$ is the j^{th} membership function of the i^{th} input vector and *n* is the number of total states. The asterisk "*" denotes a fuzzy *t*-norm operator. Here the product inference method is used as the *t*-norm operator.

Step 4:

Output Generation with Memory Weight Learning (*W*→*Y*): Due to partial proportional fuzzy rules and existent overlap situation, more than one fuzzy rule is fired simultaneously. The consequences of the multi-rules are merged by a defuzzification process. The defuzzification approach we applied is to sum assigned weights of the activated fuzzy rules on their firing strengths, denoted as $C_j(x)$. The output of the network is,

$$y = \sum_{j=1}^{N} (w_j C_j(x) / \sum_{i=1}^{N} C_i(x)) \tag{13}$$

The weight update rule for FCMAC is as follows:

$$w_j^{(i)} = w_j^{(i-1)} + \frac{\alpha}{m}(y_d - y) C_j(x) / \sum_{i=1}^{N} C_i(x) \tag{14}$$

where α is the learning rate, *m* is the size of floor (called generalization), and y_d is the desired output. Here, an adaptive learning rate is introduced [22]. Let the tracking error *e*(*t*) be

$$e(t) = y_d - y(t) \tag{15}$$

where t is the time index. A Lyapunov function can be expressed as

$$V = \frac{1}{2} e^2(t) \tag{16}$$

Thus, the change in the Lyapunov function is obtained by

$$\Delta V = V(t+1) - V(t) = \frac{1}{2}\left[e^2(t+1) - e^2(t) \right] \tag{17}$$

The error difference can be represented by

$$\Delta e(t) \approx \left[\frac{\partial e(t)}{\partial W} \right] \cdot \Delta W(t) \tag{18}$$

Using the gradient descent method, we have

$$\Delta w_j(t) = -\frac{\alpha}{m} \frac{\partial V(t)}{\partial w_j(t)} \tag{19}$$

Since

$$\frac{\partial V(t)}{\partial w_j(t)} = e(t) \cdot \frac{\partial e(t)}{\partial w_j(t)} = (y_d - y(t)) \cdot (-C(x)^T / \Sigma C(x)) \tag{20}$$

Thus

$$\Delta w_j(t) = \frac{\alpha}{m} (C_j(x) / \Sigma C(x)(y_d - y(t)),\ j = 1,2,\ldots..,N \tag{21}$$

$$\Delta W(t) = \begin{bmatrix} \Delta w_i(t) \\ \Delta w_2(t) \\ \vdots \\ \Delta w_N(t) \end{bmatrix} = \frac{\alpha}{m} \begin{bmatrix} (C_1(x) / \Sigma C(x) \\ (C_2(x) / \Sigma C(x) \\ \vdots \\ (C_N(x) / \Sigma C(x) \end{bmatrix} \cdot (y_d - y(t))$$

$$= \frac{\alpha}{m} \cdot (C(x) / \Sigma C(x) \cdot (y_d - y(t)) \tag{22}$$

From (13) and (15) we have

$$\frac{\partial e(t)}{\partial w_j(t)} = -C_j(x) / \Sigma C(x),\ \frac{\partial e(t)}{\partial W} = -C(x)^T / \Sigma C(x) \tag{23}$$

From (16) to (23) we have

$$\begin{aligned}
\Delta V &= \frac{1}{2}\left[e^2(t+1)-e^2(t)\right] \\
&= \frac{1}{2}\left[e(t+1)-e(t)\right]\cdot\left[e(t+1)+e(t)\right] \\
&= \frac{1}{2}\left[\Delta e(t)\right]\cdot\left[2\cdot e(t)+\Delta e(t)\right] \\
&= \Delta e(t)\cdot\left[e(t)+\frac{1}{2}\Delta e(t)\right] \\
&= \left[\frac{\partial e(t)}{\partial W}\cdot\frac{\alpha}{m}\cdot(C(x)/\Sigma C(x))\cdot e(t)\right]\cdot \\
&\quad \left\{e(t)+\frac{1}{2}\left[\frac{\partial e(t)}{\partial W}\cdot\frac{\alpha}{m}\cdot(C(x)/\Sigma C(x))\cdot e(t)\right]\right\}
\end{aligned} \tag{24}$$

Since $\frac{\partial e(t)}{\partial W}=-C(x)^T/\Sigma C(x)$, we have

$$\begin{aligned}
\Delta V &= \left[-C(x)^T/\Sigma C(x)\cdot\frac{\alpha}{m}\cdot C(x)/\Sigma C(x)\cdot e(t)\right]\cdot \\
&\quad \left\{e(t)+\frac{1}{2}\left[-C(x)^T/\Sigma C(x)\cdot\frac{\alpha}{m}\cdot C(x)/\Sigma C(x)\cdot e(t)\right]\right\} \\
&= \left[-\frac{1}{2}\frac{\alpha}{m}\cdot e(t)\cdot\frac{C(x)^T\cdot C(x)}{(\Sigma C(x))^2}\right]\cdot\left[2e(t)-\frac{\alpha}{m}\cdot e(t)\cdot\frac{\left\|C(x)^2\right\|}{(\Sigma C(x))^2}\right] \\
&= \left[-\frac{1}{2}\frac{\alpha}{m}\cdot e^2(t)\cdot\frac{\left\|C(x)^2\right\|}{(\Sigma C(x))^2}\right]\cdot\left[2-\frac{\alpha}{m}\cdot\frac{\left\|C(x)^2\right\|}{(\Sigma C(x))^2}\right]
\end{aligned} \tag{25}$$

Let $\left[2-\frac{\alpha}{m}\cdot\frac{\left\|C(x)^2\right\|}{(\Sigma C(x))^2}\right]>0$ then $\Delta V<0$, i.e., select the learning rate as

$$\frac{2\cdot m\cdot(\Sigma C(x))^2}{\left\|C(x)^2\right\|}>\alpha>0 \tag{26}$$

This implies that $e(t)\to 0$ for $t\to\infty$. Thus, the convergence of the adaptive FCMAC learning system is guaranteed.

5 Experiment Results

First, the path to the elevator is planned with an indoor localization system. The WMR moves along the path and reaches the desired location in front of the elevator buttons and then uses the webcam to obtain button images. After using image recognition to find location information, the WMR moves its robot arm to reach the target altitude and presses the desired button. The WMR then pulls back and checks the button light. With the indoor localization system the WMR moves along the planned route to reach the elevator door and uses ultrasonic sensors to detect the elevator door. After the elevator door opens, the WMR moves into the elevator and tracks the preset path so that it can reach the desired location in front of elevator buttons. Path tracking is shown in Fig. 9.

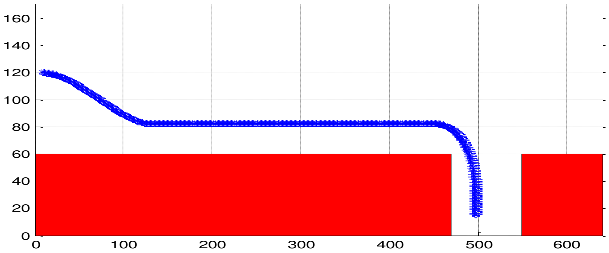


Figure 9
WMR path tracking by FCMAC

With image identification, the WMR moves its robot arm and presses the desired button, and then it waits for the elevator door to open. When the elevator door opens, the WMR moves out from the elevator, and then all actions are completed.

When taking the elevator, the WMR starts at the outside of the elevator. First, the WMR gets its own coordinates via the StarGazer. The WMR moves along the desired path to the front of the outside button of the elevator by path tracking until the WMR moves to the stop line. In the second part, the WMR raises its left arm and points to the target button. When the WMR pushes the outside button, it checks the button lights and moves backward to the stop line, as shown in Fig. 10.

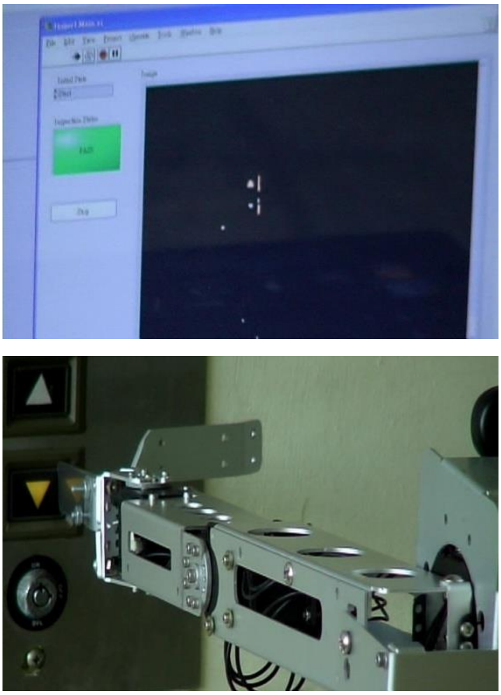


Figure 10
Push the outside button and identify it

In the third part, the WMR moves to the position in front of the elevator. This mission needs to use the coordinates from the StarGazer. With the coordinates, the WMR can move close to the desire path, as shown in Fig. 11.

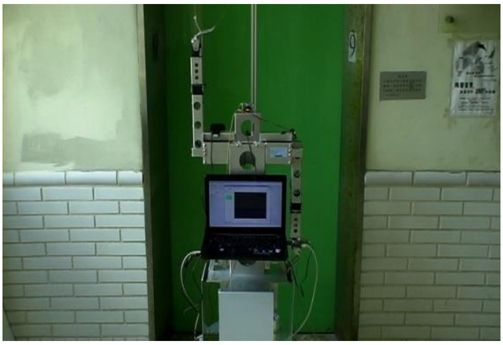
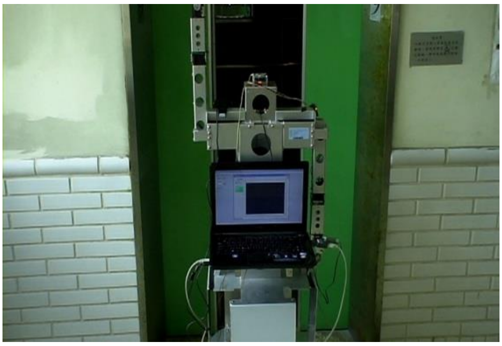


Figure 11
Move to the position in front of the elevator

In the fourth part, the WMR waits in front of the door, and when the door opens the WMR moves to the target position, which is in front of the elevator button, as shown in Fig. 12.



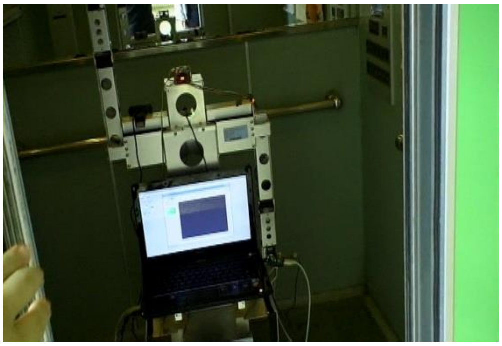


Figure 12
Wait for the door open and move into the elevator

In the fifth part, the WMR pushes the inside button of the elevator and checks the button light. First, the WMR raises its left arm and points to the target button. Second, when the WMR pushes the inside button, it identifies the button lights and move backward to the stop line, as shown in Fig. 13.

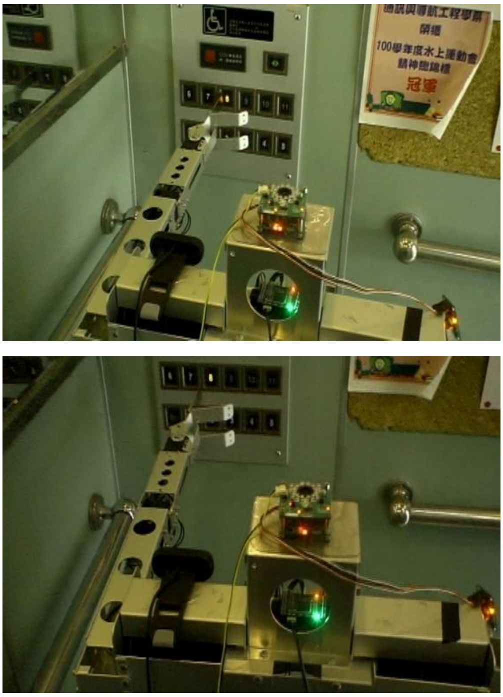
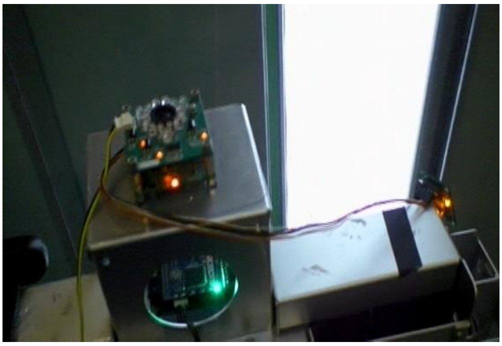


Figure 13
Push the inside elevator button and identify it

In the sixth part, the WMR moves to the position in front of the elevator door, waits for the door to open and moves out, as shown in Fig. 14. Then the complete processes are finished.



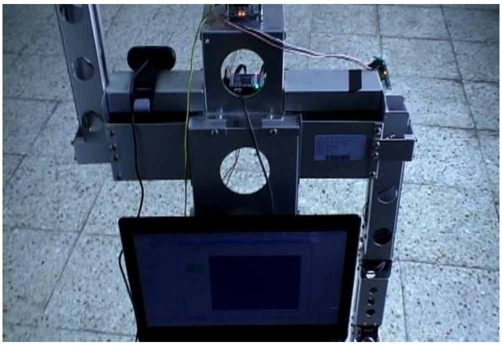


Figure 14
Move out of the elevator

Conclusions

In this paper, an intelligent control scheme based on FCMAC control, image processing, and adaptive resonance theory is proposed to control a mobile robot for path tracking and automatically taking an elevator. This research integrates a localization system, StarGazer, and ultrasonic sensors to track the desired path and distinguish whether the elevator door is opened or not. With the coordinate provided by the StarGazer and the controller, the WMR can move along the desired path. In image processing, the interference of light intensity is very troublesome. Therefore, the HSV color space is used to solve interference of light. The webcam vibrates while the WMR moves and this will cause the captured image to be unstable and with noise. This problem can be solved with filtering processes. Experiment results show the developed control system can drive the wheeled robot to track the desired path and to take the elevator successfully.

References

[1] J. J. Zhan, C. H. Wu, J. G. Juang: Application of Image Process and Distance Computation to WMR Obstacle Avoidance and Parking Control, 2010 5th IEEE Conference on Industrial Electronics and Applications, pp. 1264-1269, 2010

[2] A. Rodic, G. Mester: Sensor-based Navigation and Integrated Control of Ambient Intelligent Wheeled Robots with Tire-Ground Interaction Uncertainties, Journal of Applied Science, Acta Polytechnica Hungarica, Vol. 10, No. 3, pp. 114-133, 2013

[3] Y. Zhao and S. L. BeMent: Kinematics, Dynamics and Control of Wheeled Mobile Robots, 1992 IEEE International Conference on Robotics & Automation, 1992, pp. 91-96

[4] Y. P. Leow, K. H. Low, and W. K. Loh: Kinematic Modelling and Analysis of Mobile Robots with Omni-Directional Wheels, The Seventh International Conference on Control, Automation, Robotics and Vision, 2002, pp. 820-825

[5] Q. H. Zhong: Using Omni-Directional Mobile Robot on Map Building Application, Master Thesis, Department of Engineering Science, NCKU, ROC, 2009

[6] Y. Chung and C. Park, and F. Harashima: A Position Control Differential Drive Wheeled Mobile Robot, IEEE Transactions on Industrial Electronics, Vol. 48, No. 4, pp. 853-863, 2001

[7] E. X. Shi, W. M. Huang, and Y. Z. Ling: Fuzzy Predictive Control of Wheeled Mobile Robot Based on Multi-Sensors, The Third International Conference on Machine Learning and Cybernetics, 2004, pp. 439-443

[8] T. H. Lee, H. K. Lam, F. H. F. Leung, and P. K. S. Tam: A Practical Fuzzy Logic Controller for The Path Tracking of Wheeled Mobile Robots, IEEE Control Systems Magazine, 2003, pp. 60-65

[9] P. Bai, H. Qiao, A. Wan, and Y. Liu: Person-Tracking with Occlusion Using Appearance Filters, The 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems Beijing, China, October 9-15, 2006, pp. 1805-1810

[10] T. Darrell, G. Gordon, M. Harville, J. Woodfill: Integrated Person Tracking Using Stereo, Color, and Pattern Detection, Conference on Computer Vision and Pattern Recognition, 1998, pp. 601-609

[11] J. S. Albus: A New Approach to Manipulator Control: The Cerebellar Model Articulation Controller (CMAC)," *J. Dynamic System, Measurement and Control*, Vol. 97, No. 3, pp. 220-227, Sep. 1975

[12] C. T. Chiang, and C. S. Lin: CMAC with General Basis Functions," *Neural Netw.*, Vol. 9, No. 7, pp. 1199-1211, 1996

[13] Y. F. Peng and C. M. Lin: Intelligent Motion Control of Linear Ultrasonic Motor with H^∞ Tracking Performance," *IET Contr. Theory and Appl.*, Vol. 1, No. 1, pp. 9-17, Jan. 2007

[14] P. E. M. Almedia, and M. G. Simoes: Parametric CMAC Network Fundamentals and Applications of a Fast Convergence Neural Structure," *IEEE Trans. Industrial Application,* Vol. 39, No. 5, pp. 1551-1557, 2003

[15] C. M. Lin, L. Y. Chen, and C. H. Chen: RCMAC Hybrid Control for MIMO Uncertain Nonlinear Systems Using Sliding-Mode Technology," *IEEE Trans. Neural Network*, Vol. 18, No. 3, pp. 708-720, May 2007

[16] G. A. Capenter and S. Grossberg: Adaptive Resonance Theory, CAS/CNS Technical Report, 2009-008, 2009

[17] User's Guide, Localization system StarGazer™ for Intelligent Robots. HAGISONIC Co., LTD, 2004

[18] idRobot iHomer USER MANUA, http://www.idminer.com.tw

[19] B. Dandrea, G. Bastin, and G. Campion: Dynamic Feedback Linearization of Nonholonomic Wheeled Mobile Robots, The 1992 IEEE International Conference on Robotics & Automation, 1992, pp. 820-825

[20] C. Bunks, The RGB Color space, http://gimp-savvy.com/ BOOK/index. html?node50.html

[21] C. Bunks: The HSV Color space,
http://gimpsavvy.com/BOOK/index.html/node50.html

[22] C. M. Lin and H. Y. Li: A Novel Adaptive Wavelet Fuzzy Cerebellar Model Articulation Control System Design for Voice Coil Motors, IEEE Transactions on Industrial Electronics, Vol. 59, 2012, No. 4, pp. 2024-2033

Autonomous Hybrid Honeypot as the Future of Distributed Computer Systems Security

Peter Fanfara, Marek Dufala, Ján Radušovský

Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice
Letná 9, 04001 Košice, Slovak Republic
peter.fanfara@tuke.sk, marek.dufala@tuke.sk, jan.radusovsky@tuke.sk

Abstract: Computer security presents one of the fastest-evolving segments in the Information Technologies (IT) area. The traditional system security approach is slightly focused on defence but more attention has been drawn to aggressive forms of defence against potential attackers and intruders. The advanced decoy based technology called Honeypot is a similar form of protection against intrusion. The paper is focused mainly on the proposal of the autonomous hybrid Honeypot and its features in cooperation with the Intrusion Detection System (IDS). The weakness of the detection mechanism is a major IDS shortcoming that can be minimized by using the hybrid Honeypot technology. The proposed architecture can be used as a solution for a rapid increase a security with the autonomous behaviour model in a distributed computer system.

Keywords: Honeypot; Hybrid Honeypot; Intrusion; Intrusion Detection System; Types of Honeypots

1 Introduction

People are able to find information and send messages quickly and easily due to the rapid spread of Internet and Web technologies. However, if we do not put a sufficiently high priority on basic system security at the same time, hackers can take over computers using malicious code through the existing system vulnerabilities and program weaknesses. A major damage to most of companies and to personal property will be caused by the attackers' invasion, destruction, theft and falsification of information. Nowadays, due to these potential threats, there is a growing interest in improving information security as well as intrusion detection.

The beginnings of intrusion detection have brought some complications. There still exists a gap between the theoretical and practical level of intrusion detection. Well-established defence of a network/system is based on using a firewall and an intrusion detection system (IDS). Once the attackers are aware that the firewall has allowed an exception for the external security service, they are able to use this service to gain access to the internal servers through the firewall. Subsequently, this can result in another attack. The IDS cannot provide additional information about the detection of enemy attacks and cannot reduce losses caused by those attacks [1].

A conventional approach to the security is considerably focused on defence, but the interest is increasingly devoted to more aggressive defence forms against the potential attackers and intruders. The protection against intrusions based on the bait by using a Honeypot is an example of this form [2].

Honeypot is an advanced decoy-based technology that simulates weak points of system security and unsecured system services. The potential attackers focus on system vulnerabilities and very often attack the system weakest points, which are simulated by Honeypot. This feature represents the nature of system security. Some Honeypot solutions like Honeyd or Honeynets are already used to increase the system security [3].

The proposed client-server architecture uses a specific hybrid Honeypot that mainly consists of existing tools such as Dionaea, Sebek and Snort for rapidly increasing security in the distributed computer systems. The proposed Honeypot has an autonomous feature that enables its use in a random deployment environment. This Honeypot will auto-configure itself on the basis of system parameters obtained via a passive fingerprint method.

The following chapters describe system security using IDS with the detection mechanism based on the advanced Honeypot technology.

2 Intrusion Detection System

The IDS can be defined as a tool or software application that monitors the activities of the computer system and/or network due to the potential occurrence of malicious activities or breaches of security policy. The IDS produces reports for the control station. It is primarily focused on identifying and recording information about any events as well as reporting similar attempts [4, 5].

2.1 Classification of Intrusion Detection System

In view of the various environment applications, the IDS can be classified into two general types [1]:

- *Host-based* – this consists of an agent located on host computer that is used for the continuous monitoring of information from the system audit data or network activities logs. This IDS sensor type typically includes a software agent. If there are unusual circumstances, the system automatically generates and sends a warning.

- *Network-based* – this is an independent platform for intrusions identification using direct capturing of transmitted network packets and monitoring several computers. Detection sensors are placed in network bottlenecks for capturing all network traffic and analyzing individual packet contents looking for dangerous operations.

On the basis of the detection method, the IDS can be divided into three types below [1]:

- *Anomaly detection* – refers to the pattern found in the data set that is inconsistent with normal behaviour. The anomaly detection provides basic performance for normal network traffic. An alarm sounds only if the current network traffic is above or below standard parameters.
- *Misuse detection* – collects previous hacker attack characteristics and patterns, which are then saved to knowledge attack database. Consequently, it can identify attacks with the same patterns and characteristics as those of previously stored attacks. The IDS cannot trigger an alarm if the hacker uses a new attack method that has not been previously reported or detected.
- *Hybrid mode detection* – represents attack detection using a combination of previous two types, resulting in a reduction of false alarms.

2.2 IDS Structure and Architecture

The IDS consists of several elements illustrated in Figure 1 where the main element is a sensor, the mechanism for analysis, responsible for intrusion detection. This sensor contains a mechanism that makes decisions regarding a breach. The sensor receives data from three main sources of information: the IDS knowledge database, system logs and audit trails. System logs may include for example file system configuration and user permissions. This information forms the basis for further decision on intrusion detection.

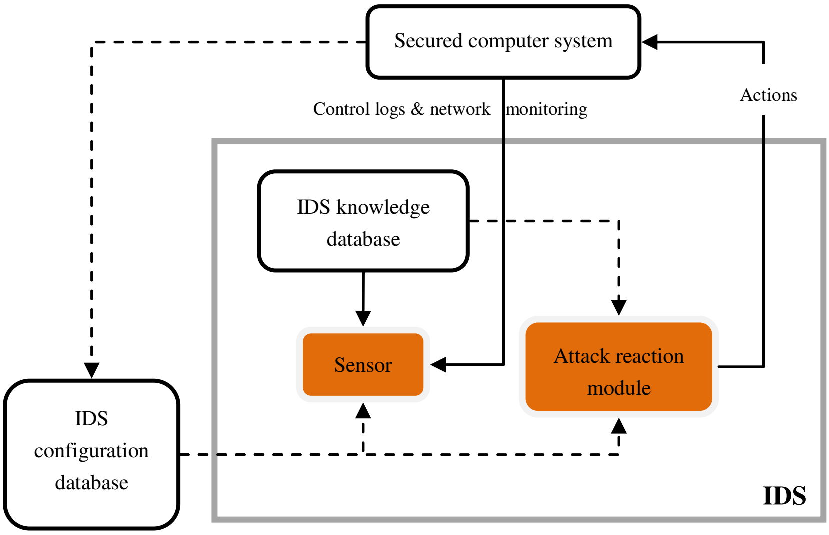


Figure 1
Intrusion detection system structure [6]

The sensor in the IDS elements illustrated in Figure 2 is integrated together with the component that is responsible for data collection, the events generator. The data collecting method is set by the policy of the events generator, which defines the filtering method for events information notifications. The events generator (operating system, network & application) in accordance with security policies produces sets of events (system logs, control records or network packets). These occurrences may be stored together with information policy either in a protected system or outside it. In some cases they are not stored, e.g. when events streams are directly transmitted to the analyzer, especially network packets [7, 8].

The role of the sensor is to filter information and to discard any irrelevant data obtained from the event file related to the protected system and to detect suspicious activity. For this purpose, the sensor uses the detection policy database, which is composed of the following parts: pattern attack, normal behaviour, profiles and necessary parameters. The database contains the IDS configuration parameters and communicating method with the reaction module. The sensor has a custom database that also includes a dynamic history of potential intrusions [7].

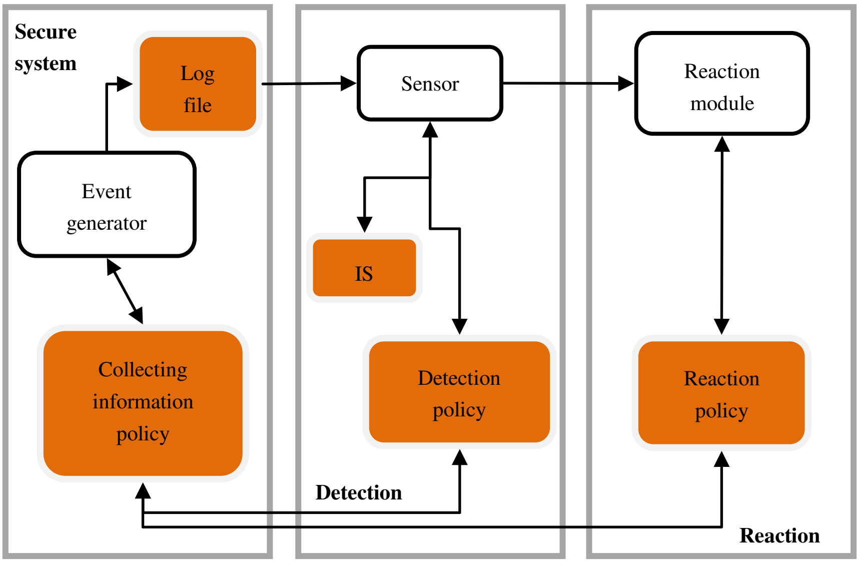


Figure 2
Intrusion detection system elements [1]

2.3 Intrusion Detection Tools

Nowadays, many IDS exist (i.e. Snort [9], SAX2, etc.), and all are specific to the system of deployment. A Snort is the most commonly used tool, one that has excellent additional conditions for usage in order to enhance the distributed system security in combination with Honeypots.

Snort represents an open-source IDS that can detect and warn of attacks (e.g. against the Honeypot). It can also capture packets and network load given by packets included in the attack. The collected information may be critical for analyzing the attacker's activities. Snort uses a modular architecture and rules based language. It combines the abnormal behaviour, detection signature and different protocol detection methods [10].

The methodology of lying and cheating by providing the emulation of some system services was domesticated in order to be able successfully monitor hackers' activities in distributed computer systems. At first sight this system appears to be legitimate. It is possible to record and monitor all hackers' activities due to the penetration and clarification of the various attackers' tactics. This idea was developed by using an advanced security tool called Honeypot.

3 Honeypot

Honeypot is a closely monitored network decoy available in different shapes and sizes, serving various purposes. It can be placed in a computer network with the firewall, in front of it and/or behind it. These points of deployment are the most frequent sites for attackers obtaining access to the system. These sites provide the best solution for acquiring the maximum amount of information about the attackers' activities. The main aim of Honeypot is to collect information by compromising the system data in the way that any system infiltration would be unfeasible to do in the future.

The main benefit of Honeypot is detection. It can address IDS shortcomings, by minimizing the amount of false positive and false negative alerts generated. There are several situations in which IDS cannot generate a warning of attack: if the attack is too short or if the appropriate security rule refers too many false alarms or detects excessive network traffic and thus drops packets. One solution is to use Honeypot, since it has no way to affect system functions. Honeypot implementation uses an unused IP address, which means that all incoming communication is almost certainly unauthorized, i.e. there are no false positive or false negative alarm warnings or large data files to be analyzed [6].

The data obtained from the Honeypots can be used to create better protection and countermeasures or system reconfiguration against future threats.

3.1 Types of Honeypots

Honeypots can be classified in different ways. Classification according to purpose and level of interaction is the most frequent one.

3.1.1 Purpose Honeypots

This basic classification divides Honeypots based on the area of deployment.

- Research Honeypot – this type is used merely for research. The main objective is to obtain as much information as possible about an intruder in a way that allows full infiltration and penetration of security system. It is used to obtain information and detect new methods and types of tools used to attack other systems as well as to analyze the hacker's traces, their identity or modus operandi. Another option in research is that the Honeypot can be used to discover potential risks and information vulnerabilities in enterprise systems [11].

  The primary function is to examine how attackers proceed and lead their attacks, which usually means understanding their motives, behaviour and organization. Research Honeypots are complex in terms of deployment, maintenance and the capturing huge amounts of data. On the other hand, they are highly useful security tools in the field of development and in enhancing forensic analysis capabilities.

In addition to the information obtained from research, the Honeypot can be used to improve prevention against attack. By improving the detection and response to attacks this Honeypot type contributes to direct security only by a small amount [12].

- Production Honeypot – it is used in organizations for protection and to help to reduce level of risk. It provides immediate enhancing of the system security [3]. Since it does not require as much functionality as the research Honeypot, its development and deployment is usually much easier. Nevertheless, it can identify various attack methods. The production Honeypot provides less information about the attacker than the research one. It is possible to determine where the attackers come from and what specific actions was performed, but it cannot determine the intruders' identities, how they are organized or which tools were used.

  The production Honeypot has minimum value as a prevention mechanism. The best way to implement this Honeypot is to use well-firewalled system, an IDS, and mechanisms for locking and fixing the system [12].

3.1.2 Level of Interaction

All Honeypots are based on the same concept: nobody should interact with Honeypot. The level of interaction can be defined as a maximum range of options available to attack allowed by Honeypot. Therefore, any transactions or interactions based on definition become illegitimate. Honeypots can also be categorized according to the level of interaction between intruders and the system. This classification helps in choosing the correct type for deploying in system [12].

- Low-interaction – does not contain any operating system (OS) for communication with the attacker. All tools are installed purely for emulation of OS and services that cannot be used to gain full access to the Honeypot. Emulation is set up to cooperate with the attacker and malicious code, resulting in radical risk reduction. Attackers can only scan the Honeypot and connect to several ports. Low-interaction Honeypots are characterized by the possibility of easy deployment and maintenance. Honeyd is an example of a low-interaction Honeypot.
- Medium-interaction – this type is more sophisticated than the previous one but still does not have installed any OS. The medium-interaction Honeypot only provides an illusion of real OS to the attacker because it contains a number of emulated services the attacker can interact with. This type is able to detect automated attacks and extract information about malware binaries. Malicious software can be automatically downloaded and analyzed. The Dionaea tool and Honeytrap are the examples of this Honeypot type.

- High-interaction – the most advanced Honeypot. On the other hand, it represents the most complex and time-consuming design with the highest rate of risk, because it implies the functional OS. It gives the attacker the ability to communicate with the real OS where nothing is simulated, emulated or restricted. This Honeypot allows for collecting the highest amount of information because it can detect and analyze all performed activities. Main focus is set to obtain valuable information about intruders by making available the entire system or even allow handling with it.

3.2 Architecture of the Hybrid Honeypot

The Hybrid Honeypot represents a combination of two Honeypots with different levels of interaction. The combination is a secure solution because it is possible to take advantage of both Honeypot types, which complement each other and thus limit their disadvantages, shown in Table 1. The ideal solution is to use a low-interaction Honeypot with a high-interaction one. The low-interaction Honeypot acts as a lightweight proxy, which relieves the high-interaction Honeypot and allows focusing on processing all IP address space network traffic [3].

Table 1
The essence of hybrid Honeypot

Low-interaction Honeypot	**High-interaction Honeypot**	**Hybrid Honeypot**
+ fast	- slow	+ fast
- no possibility to detect unknown attack	+ possibility to detect unknown attack + 0 false produced warnings	+ possibility to detect unknown attack + 0 false produced warnings
+ resists to time-bomb + handles interaction with attackers	- unable to resist time-bomb and can't handle interaction with attackers	+ resists to time-bomb + handles interaction with attackers
+ cheap	- expensive	+ relatively expensive
+ simple to set up and maintain	- complicated to set up and maintain	- complicated to set up and maintain

It is impossible for each proposed Honeypot not to use the implementation tools that have considerable importance in improving system security.

Dionaea is a modular architecture using a low-interaction Honeypot. It is able to simulate the server's main services and vulnerabilities due to attracting attacker/attack attention or the withdrawal of the malicious code [9].

Sebek is the most advanced tool for comprehensive data collection, aiming to capture as much information about the attackers' activities as possible from the Honeypot by stopping specific system calls (*syscalls*) on the kernel level [13].

3.3 Advantages and Disadvantages

All security technologies have some risk margin. If knowledge and experience represent the power of attackers, they also provide advantages for security professionals. By knowing the Honeypot risks, it is possible to use knowledge to mitigate them and reduce the disadvantages [11].

Honeypots have several unique advantages that are unique to this advanced technology [3]:

- Small data sets – Honeypots can monitor only the traffic that comes directly to them. They collect small amounts of data, but on the other hand, they may contain high value information.
- Minimal resources – Honeypots require minimum system resources for capturing harmful activities. Systems with low-end specifications will be enough to run a Honeypot.
- Discovery of new tools & tactics – Honeypots capture everything that starts interactions with them.
- Encryption or IPv6 – Honeypots can also operate in encrypted or IPv6 environments/systems.
- Simplicity – Honeypots are very easy and flexible to operate, so they do not need complicated algorithms to function properly.

The decoy-based technology, like other security solutions, also has its own disadvantages, which are described bellow [3]:

- Risk of takeover – if an attacker takes control of the Honeypot, he can exploit it to attack other systems inside or outside the system of deployment.
- Limited vision – Honeypots can only monitor the traffic that comes directly to them.
- Discovery and fingerprinting – Honeypots have some expected characteristics or behaviours. If the attacker uses some fingerprinting tool, he can identify the working Honeypot in attacked system. Even a simple error, such as a misspelled word in the emulated service, can act as a Honeypot signature.

4 IDS Architecture Using a Sophisticated Hybrid Honeypot

The main IDS weakness lies in the ability to detect new attack types. The use of different attack strategies or new tools cannot be detected by IDS. These new attacks need to be registered in the IDS configuration database and only then is it possible to detect them. The proposed IDS uses a hybrid Honeypot with an autonomous ability to reduce the risk of detection failure, and it provides extensive data collection. The Hybrid Honeypot also affords the opportunity to design safety features of distributed systems through the captured data. It also minimizes any system intrusion threats. Hybrid Honeypot combines several tools: Snort, Dionaea and Sebek. The proposed system, illustrated in Figure 3, analyzes all captured various data formats due to the rapid response to attacks. It also serves as a warning reporting system to the system administrator via web interface when interaction with Honeypot occurs.

The proposed intrusion detection system contains existing client-server detection architecture and its arrangements for using the proposed sophisticated Honeypot technology.

The architecture consists of several clients and a central server. Clients collect information about an attack and the captured malware is sent back to server. The server records and analyzes the received data, issues a warning and displays the overall information to the system administrator via web interface. The architecture is designed to achieve the effect of centralized distributed information management and to build complex distributed system of early warning for distributed computer systems.

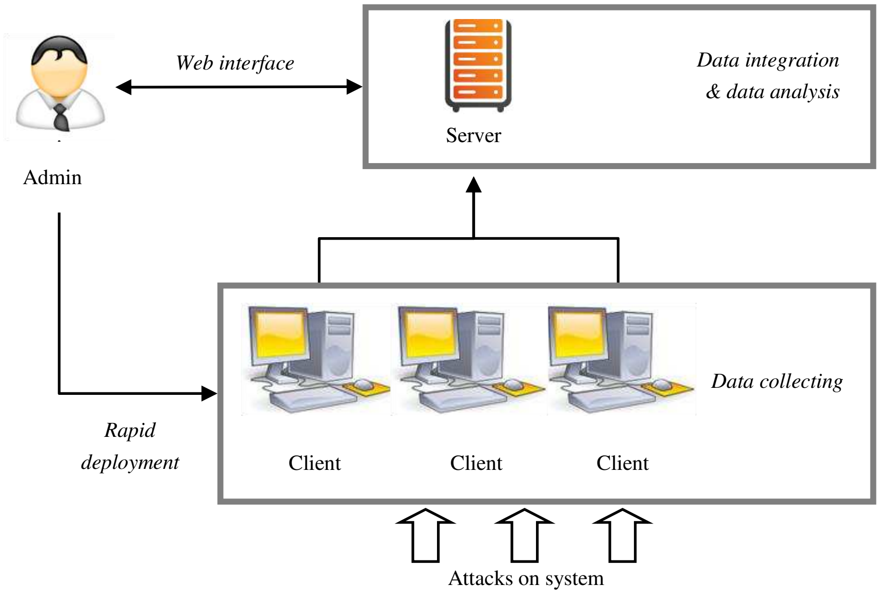


Figure 3
Architecture of proposed detection system

4.1 Client Architecture

The clients are installed in the same domain because of the data-gathering activities during the attack. Diverse system components for collecting data sets are activated depending on the different type of cyber-warfare activities. Then the data sets are sent to server for further analysis and they subsequently update the system security. The client architecture, shown in Figure 4, consists of three components:

- Snort – monitors and filters packets during intrusion detection. It identifies the patterns and characteristics of specific attacks, information and warning messages.
- Dionaea client – simulates general services and vulnerabilities that attract the attackers. It captures malware patterns and characteristics.
- Sebek client – records the attacker behaviour during interaction with Honeypot into the log files.

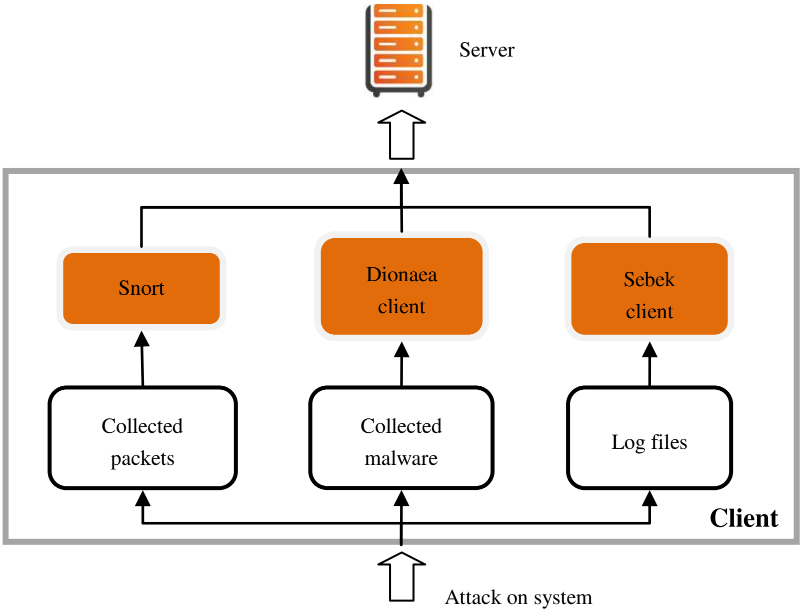


Figure 4
Client architecture

4.2 Server Architectur

At the same time, the server is connected to multiple clients owing to centralization of the collected data, and it is set to receive all outgoing messages, which are then stored in the database. The server architecture is shown in Fig. 5.

It indicates that the attacker's intention is targeted to extensive computer or scanning attacks by using individual interconnection reports. The architecture of the server consists of three main parts, the outputs that are normalized before they are stored into the database:

- Dionaea server – receives malware patterns sent by the Dionaea client component.
- Sebek server – simultaneously receives and filters multiple data sources representing the instruction or cohesion of data sent to be stored.
- Verification – modular design of open-source hybrid system for detecting an intrusion using standard communication format. The verification part can receive the data from many clients and integrate disparate data formats.

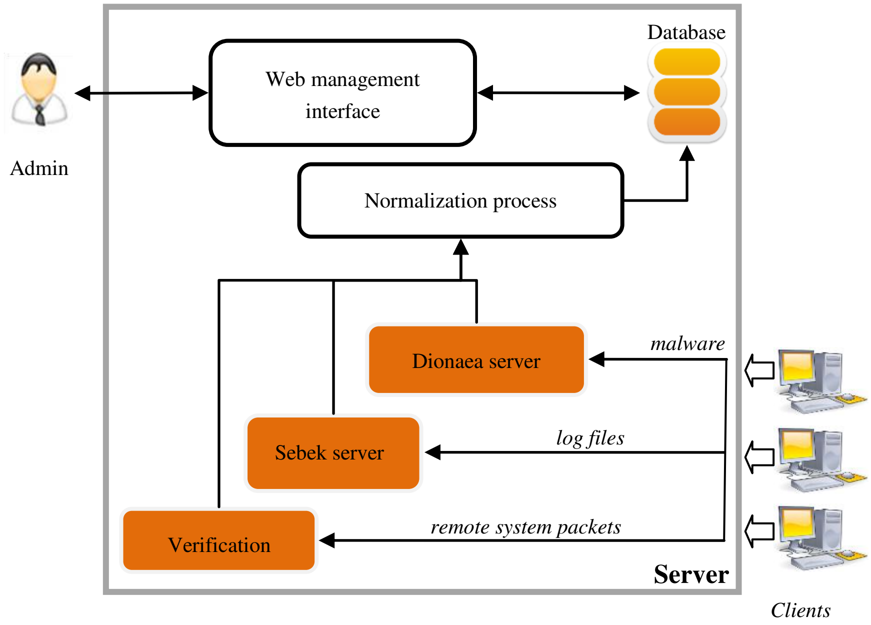


Figure 5
Server architecture

Web server interface displays all attack analyses obtained from the database. At the same time it monitors the attack patterns and occurrences of unusual circumstances. In the event of their occurrence, the specific messages are highlighted via web management interface for the correct and in time response.

4.3 Sophisticated Hybrid Honeypot

4.3.1 Desired State

The sophisticated hybrid Honeypot is a suggestion to address the current state of the security system that works on the plug-&-play principle. The optimum state will occur when the Honeypot runs a complete configuration process right after plug-in. For example, after installing the Linux OS to the distributed computer system, we will have the Linux Honeypot, or after removing any of services, the related service will be removed from the Honeypot list of emulated services. When replacing routers in the system, for example changing routers from Hewlett-Packard vendor to Cisco routers, the existing Honeypot, pretending to be a router, will immediately and autonomously auto-reconfigure and update itself. The solution is a device that simply connects to the network/system and learns the topology autonomously without any external support. Upon completion of the scanning process, the device will accurately determine the number of Honeypots with their configuration and it will be able to adapt quickly to any modifications in the system.

4.3.2 Trouble

The most critical component of the sophisticated Honeypot is the method how the Honeypot gets the information about the deployment network, for example, what systems are used and how they are used in the current environment. An example of heterogeneous distributed computer system is shown in Figure 6. The Honeypot will be able to sophistically map and promptly respond to the current system environment after obtaining the network parameters. One of the simplest possible ways is an active probing and thus determining the system and type of used services. The use of the active method of data mining also has some shortcomings in terms of increased network load; there is a risk to the running system functionality.

The sophisticated Honeypot would have to constantly scan all active environments of deployment to remedy the described lack. This solution is not the most appropriate approach.

4.3.3 Solution

The solution to the drawbacks of active system scanning is a passive approach, specifically the passive fingerprinting and mapping method. The passive fingerprinting method is not new. The idea is to obtain the system overview via mapping the current environment. The difference vis-a-vis the active method is that it has a different mapping approach. This approach is based on obtaining information through passively capturing network traffic, analyzing it and then determining the system identity based on the unique system fingerprints. The passive method uses the same method as the active one but in different ways.

Tools, such as Nmap [14], create a signature database that contains known operating systems and services. All searching tools actively broadcast packets that require a response from destination devices right after creating a signature database. Incoming responses are unique to most operating systems and services. Responses are simply compared with the data in the signature database due to a clear identification of the operating system and used services.

Passive fingerprinting uses the same approach as the signature database, unless the data are obtained passively. Instead of actively probing the system, the passive fingerprinting method intercepts network traffic and analyzes the captured packets, which are then compared with a signature database. After the analyzing process ends, the concrete operating system should be known. Passive fingerprinting is not limited to use only with the TCP protocol, which allows for the use of other protocols. The usage of the passive method represents several advantages: less likelihood of damage or shutting down of the system or service, and the ability to identify systems using a firewall. The passive method is continuous, which means changes in the network structure are captured in real time [8]. This advantage becomes a critical feature in maintaining a realistic Honeypot over a longer time period. The only disadvantage of the passive method is the correctness of functioning through the routed networks; the most effective usage of passive obtaining parameters method is in local area networks.

4.3.4 Concept

The proposed Honeypot data obtaining mechanism is based on the concept of the passive fingerprinting method. The Honeypot is deployed as an independent device that is physically connected to the computer network of a distributed computer system. The tracking and learning phase starts after connecting to a network device. In this phase, the Honeypot learns the topology and plans the deployment of other Honeypots. The duration of learning phase is variable and depends on the system topology. The proposed Honeypot can determine the number of used systems, the types of operating systems, and the running services via passive analyzing of the network traffic. It also has the feature to determine with whom and how often a concrete system or service communicates. This information is used for mapping and obtaining knowledge about the deployment network.

Once the Honeypot collects all the necessary information, it can start with the Honeypot deployment illustrated in Figure 7. The created Honeypots are designed to mirror the real system and decrease the risk of the attack. Honeypots with the ability to look and behave in the same way as the production environment can easily blend with their surroundings. Their identification and tracing by attackers is much more difficult.

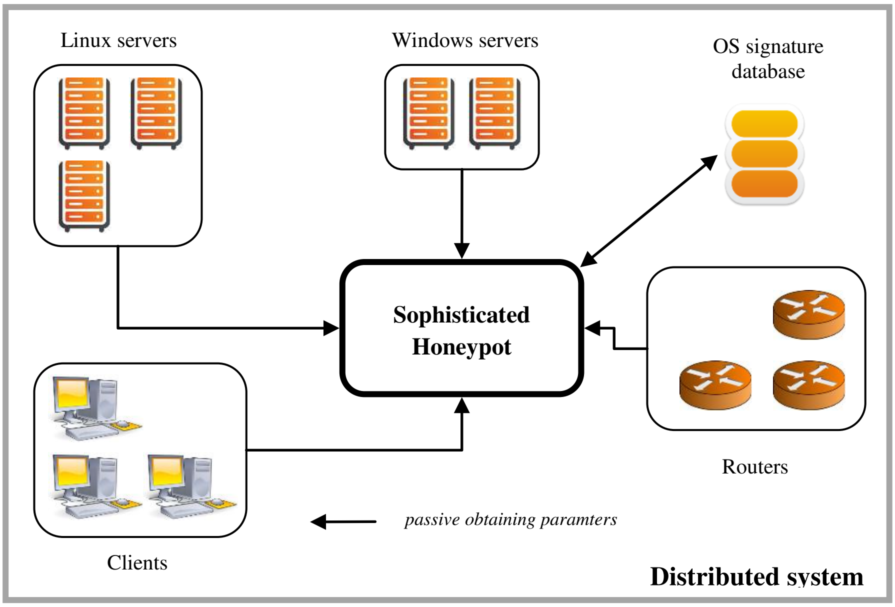


Figure 6
Passive obtaining system parameters via sophisticated Honeypot during determination process of deployment virtual Honeypots

Passive acquisition of information does not end but rather is continuous. It monitors the entire network system and increases its flexibility. Any change is identified in real time and the necessary steps (the system deployed Honeypots) are realized in the fastest possible time.

The proposed sophisticated Honeypot considerably reduces the work necessitated by configuration and administration in a constantly changing environment.

4.3.5 Deploying Honeypots in a System

The traditional solution to the issue of implementing the Honeypot in the system requires the physical placement of a new computer for each monitored IP address. The physical Honeypot deployment represents considerable time and work. An autonomous and simpler solution, for example, fire-&-forget, is not to implement a physical Honeypot but rather a virtual type, which, if in sufficient quantity, can monitor all the unused IP addresses. Virtual decoys pursue identical IP address space as the system itself. All virtual decoys are designed, located and managed by only one physical device, the proposed sophisticated Honeypot illustrated in Figure 7.

Whereas virtual Honeypots monitor unused IP addresses in computer networks, it is almost certain that any activity detected on the monitored IP addresses is most likely a malicious or unauthorized behaviour. Using information gathered through passive mapping of the environment can determine the quantity and type of Honeypot deployment.

The ability to dynamically create and deploy virtual decoys already exists. An open-source solution with a low interaction Honeypot called Honeyd [15] allows for deploying virtual environment decoys throughout the organization. It is possible to realize the design of a sophisticated autonomous Honeypot with dynamic creation and deployment of virtual decoys that minimizes the risk of detection and identification of intruders by merging the surrounding environment with a combination of options, such a Honeyd solutions capabilities and passive fingerprinting.

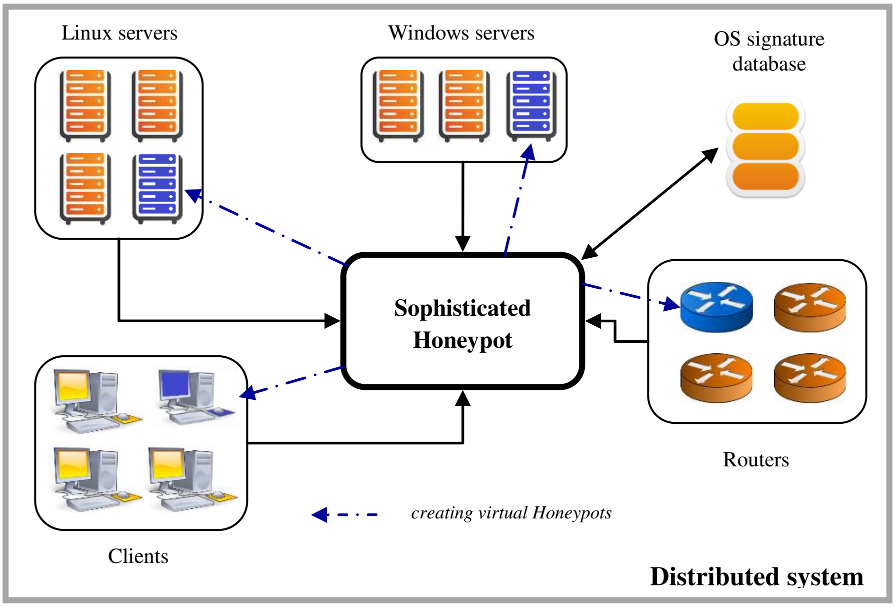


Figure 7
Deployment of virtual Honeypots based on obtained parameters

Conclusion

The security of information technologies is essential in a society that depends on information. Therefore, considerable emphasis is placed on data and information source protection in the systems development processes [16, 17]. The protection of access, availability and data integrity represents the basic safety features required for information resources. Any disruption of these properties will end in penetration into the system and would increase security risk. One way of defense is a system that detects unusual and suspicious behavior, called the IDS. IDS major risk is represented by undetected penetration problem.

An advanced technology called Honeypot has huge potential for the security community and it can also achieve several objectives of other technologies, which makes it almost a universal solution. The usage of Honeypots represents a cost-effective answer to improvements in the organization security status.

Honeypots, like any new technology, also have some shortcomings that need to be overcome and removed. The use of decoy-based technology represents a cost-effective solution to increase the security status of the organization. Therefore, they are being deployed in systems at an increasing rate, but mostly as a passive device. Many system administrators monitor the situation in the system via Honeypot, and if a production environment is attacked once, administrators analyze and implement solutions manually; the Honeypots' capabilities are not used at all, or they are used minimally. Despite the many advantages of Honeypot, it is not a panacea for all breaches of system security. Since it is used for gathering information about attacker and other threats, it is useful as an IDS detection mechanism.

The future of Honeypots and cyber security intrusion detection lies in sophisticated (autonomous hybrid) decoys. They have a radical revolutionary assumption in autonomous deployment and maintenance. They are becoming a highly-scalable solution due to their capability to study and monitor the network real time. Deployment and management becomes more cost-effective and also provides better integration into the system of deployment. Another advantage of the proposed Honeypot lies in minimizing the risk of human errors during manual configuration. The merger surrounding environment also minimizes the risk of being identified by the attacker.

Acknowledgement

This work has been supported by the Slovak Research and Development Agency under the contract No. APVV-0008-10 (20%). This work is also a result of implementing the Project: Development of the Centre of Information and Communication Technologies for Knowledge Systems (ITMS project code: 26220120030) supported by the Research & Development Operational Program funded by the ERDF (80%).

References

[1] J. McHugh, A. Christie, J. Allen: "Defending Yourself: The Role of Intrusion Detection System," IEEE Software, IEEE Computer Society, pp. 42-51, October 2000

[2] F. G. Lyon: "Nmap Network Scanning: The Official Nmap Project Guide to Network Discovery and Security Scanning," [online], Nmap Project, USA, ISBN 978-0979958717, January 2009. Available on: <http://nmap.org/book>

[3] S. Karthik, B. Samudrala, A. T. Yang: "Design of Network Security Projects Using Honeypots," Journal of Computing Sciences in Colleges, 2004

[4] L. Vokorokos, N. Ádám, A. Baláž: “Application of Intrusion Detection Systems in Distributed Computer Systems and Dynamic Networks,” Computer Science and Technology Research Survey, Košice, 2008, pp. 19-24, ISBN 978-80-8086-100-1

[5] L. Vokorokos, N. Ádám, A. Baláž, J. Perháč: “High-performance Intrusion Detection System for Security Threats Identification in Computer Network,” Computer Science and Technology Research Survey, Košice, 2009, pp. 54-61, ISBN 978-80-8086-131-5

[6] R. Baumann, C. Plattner: “White Paper: Honeypots,” Swiss Federal Institute of Technology, Zurich, 2002

[7] L. Vokorokos et al: “Architecture of Intrusion Detection System Based on Partially Ordered Events“, Computer Science and Technology Research Survey, Vol. 2, 2007, pp. 80-91, ISBN 9788080860714

[8] L. Vokorokos, A. Pekár, N. Ádám: “Data Preprocessing for Efficient Evaluation of Network Traffic Parameters,” INES 2012: IEEE 16th International Conference on Intelligent Engineering Systems, 2012, Lisbon, Portugal, pp. 363-367, ISBN 978-1-4673-2695-7

[9] Snort [online]. Available on: <http://www.snort.org>

[10] M. Tomášek, M. Čajkovský, B. Madoš: “Intrusion Detection System Based on System Behavior”, SAMI 2012: 10th IEEE Jubilee International Symposium on Applied Machine Intelligence and Informatics: proceedings: Herľany, Slovakia, 2012, pp. 271-275, ISBN 978-1-4577-0195-5

[11] L. Spitzner: “The Value of Honeypots, Part One: Definitions and Values of Honeypots,” Security Focus, 2001

[12] L. Spitzner: “Honeypots: Tracking Hackers,” Boston, USA: Addison-Wesley, Pearson Education, 2003, ISBN 0-321-10895-7

[13] Dionaea catches bug [online]. Available on: <http://dionaea.carnivore.it/>

[14] R. Chandran, S. Pakala: “Simulating Network with Honeyd,” [online], Technical paper, Paladion Networks, December 2003. Available on: <http://www.paladion.net/papers/simulating_networks_with_honeyd.pdf>

[15] N. Provos: “Developments of the Honeyd Virtual Honeypot,” [online]. Available on: <http://www.honeyd.org>

[16] Cs. Szabó, L. Samuelis: “Notes on the Software Evolution within Test Plans,” Acta Electrotechnica et Informatica, Vol. 8, Issue 2, pp. 56-63, 2008, ISSN 1335-8243

[17] Cs. Szabó, L. Samuelis: “The A-shaped Model of Software Life Cycle,” SAMI 2007, Slovakia, pp. 129-135, ISBN 9789637154560

[18] Sebek [online]. Available on: <http://www.honeynet.org/tools/sebek/>

[19] P. Jakubčo, E. Danková: “Distributed Emulation Using GPGPU,” Electrical Engineering and Informatics 2: Proceeding of the Faculty of Electrical Engineering and Informatics of the Technical University of Košice, Slovakia 2011, pp. 284-287, ISBN 978-80-553-0611-7

The Cxnet Complex Network Analyser Software

Árpád Horváth

Óbuda University, Alba Regia University Centre, Székesfehérvár, Hungary
horvath.arpad@arek.uni-obuda.hu

Abstract: The study of complex networks has become important in several fields of science such as biology, sociology and physics. The collection of network data and the storage, analysis and visualisation of these data have become important contributors to the knowledge of programmers working in these fields. Our cxnet software connects several software packages of the Python language to make these tasks easier. One of the main goals of this development is to provide a comfortable application programming interface for students to develop their own programs. The cxnet software package is able to create the software package network of the Ubuntu Linux distribution. This network is a directed network with several types of vertices and connections. It changes quite fast and can be created easily. These properties make it an ideal object of investigation. The present paper describes some useful measures of the properties of the complex networks, the usage of the cxnet package with some examples, and our experiences in the education.

Keywords: complex network; graph theory; education; software

1 Introduction

There are complex systems that can be described as many objects with connections among them. These systems are called *complex networks*, the objects are called *vertices* or *nodes*, and the connections are called *edges*. Most of the networks are growing: many new vertices appear and some vertices disappear, and it is also true to the edges. These networks can be described mathematically as a series of graphs, so the literature of complex networks uses almost the same terminology as graph theory. Networks can be *undirected*, if the vertices in both ends of the edges have the same role, like in the Internet, or *directed*, like the Worldwide Web, where the pages are the vertices and the edges are the links between the web pages. The *degree* of a vertex refers to how many edges join to it. One of the most important properties of networks is the degrees of its nodes, which can be described and plotted as a degree distribution.

The computers and the Internet made it possible at the end of the 20th Century to collect data on large networks, and it turned out that most networks have a power-law function as their degree distributions; these types of networks are called *scale-*

free networks [1]. The average degree is not a typical degree in these networks. In scale-free networks there are a lot of vertices with a small number of neighbours and there are some nodes, the *hubs*, with a number of neighbours a magnitude bigger than the average degree (the arithmetic mean of the degrees). The degree distribution is just one of the many useful measures that can be used to describe a complex network.

Our aim was to develop software that matches the demands of teaching complex networks in higher education, software that is easy to use and extend, open source, and possibly free of charge. There are many tools that can be used to investigate networks. There are network analyser programs with graphical user interface, such as Network Workbench [2] and Pajek [3], which can be used for graph visualization and analysis. They are easy to use, but developing new algorithms for them is quite difficult. Network Workbench runs on most operating systems that support Java. Pajek runs on Windows and with Wine on Linux. The Boost Graph Library implements many graph algorithms in C++ [4], but programming in C++ needs quite a lot of skills. As we stated, Python language is object oriented, but it has a very clear syntax that can be learned with middle-level programming skills, and we decided to use that language. In the Python language, there are two notable network analyzer packages: the igraph and NetworkX. These packages have implemented the most important algorithms of network analysis, but – as their developers do not want them to depend on many other packages – they do not have functions for plotting degree distribution and other functions that are useful in visualizing network properties and for acquisition of the data of special networks.

We developed the cxnet package of the Python language to make the investigations of networks easier. The Python language is ideal for education because it has a straightforward syntax, an excellent interactive shell and many useful standard and third party packages, and it is a free software. The cxnet package has many features beyond the features of the packages it is based on. It can create the network of the software packages of the Debian and Ubuntu Linux in tens of seconds and store these networks into files for further investigation. It can create and plot the degree distributions of any network, and plot it with several binning methods or as cumulative distribution. In the event the distribution is scale-free, it can determine its exponent with an equation that can be derived by the maximum likelihood method and plot a power-law function with this exponent easily. It is able to plot the neighbourhood of a package with edges and vertices coloured according to type, to list the names of the packages with the highest degree, indegree or outdegree, and to create statistics about the types of the vertices and edges. All but plotting can run on a server computer without graphical user interface, so it can be used to automate data acquisition and in dynamic web pages as well.

The four Python packages cxnet is based on to achieve these goals are:

- *igraph* for investigating, plotting, storing and loading networks,
- *matplotlib* for plotting diagrams, and
- *python-apt* for obtaining the properties of the Linux packages and the connections among them.

In our centre, the students may participate in a course called “Investigation of complex networks”. This gives students the possibility to learn the theoretical basics of network science: the most important measures characterizing the structure of the networks, the models of growing networks, and methods to generate networks with given properties. They learn not only the theory, but also how to analyse networks and write programs using the software packages of the Python language, including the *cxnet* software package. In addition to the network of the Linux software packages, the cxnet package can fetch a lot of archived networks to study from its webpage.

2 Requirements and Installation of the Cxnet Software Package

The cxnet software package is written in the Python language, a general purpose object oriented language [5]. This package is based on several other packages. The most important dependency is the *igraph* network analyser package [6]. It is written mainly in C language to reach its goal, to be able to analyse huge networks fast. It can be used not only in C and C++ programs; it also has extensions for the Python and R languages. The igraph package has implemented most of the algorithms of the complex network science as well as an advanced network visualisation. Earlier version of cxnet based on another network analyzer package was named NetworkX [7]. To plot functions such as the degree distribution we need the *pylab* Python module of the *matplotlib* plotting package [8]. Pylab uses the *array* data type of the *numpy* package to store row vectors of float numbers and the *matplotlib* module to plot functions. Pylab combines these two modules to get functionalities nearly equivalent to that of MATLAB. The aim of our package is not to hide the possibilities of igraph, pylab and numpy, but to help to use them together. The last package is the python-apt [9], which is needed to provide cxnet with information about the software packages of Linux.

These packages can be installed on Windows, Linux and Mac OS X as well, except for the python-apt package. We can create the software package network just on Linux distributions using the APT package management tool, such as Ubuntu and Debian. APT is the standard installation tool of Ubuntu and its detailed description can be found in the next section. It is recommended to install cxnet on Ubuntu Linux distribution version 12.04 or later. On the Ubuntu 12.04, the numpy, python-apt, ipython, igraph, networkx and matplotlib software packages can be installed using the APT package management tool. The stable and development version of the cxnet package can be downloaded by the git version control system. The up-to-date description of the installation process can be found in the documentation of cxnet [10].

The cxnet package is a rather complex program. Large part of the package can be tested automatically with the unit tests included in the package. Running of these unit tests (and the unit tests of the igraph package) is advisable to test that the package works on that platform as expected.

The cxnet and igraph packages use the unittest module of the standard library of the Python program language for unit tests. This makes possible to check, that the functions returns with the expected value for the given parameters, and raises the expected exceptions for the parameters, that have no sense.

3 The Software Package Network of Linux

Most of the Linux distributions have precompiled binary files for one or more processor architectures. Some distributions, like Ubuntu and Debian, use the same package management format, the *deb* format, developed by the Debian developers. The most convenient way to install these packages is to use the APT package management tool in command line or graphical user interface. APT can handle package repositories stored on the Internet. APT needs to know the dependencies of the packages to install, update or remove them. These dependencies and other connections are stored in compressed text files that can be downloaded from the repositories. In these files the dependencies (the packages a package depends on) of all packages are listed. There are, however, dependencies that are not *real packages*, and they cannot be downloaded; these packages are called *virtual packages*. In Figure 1, the editor package is a virtual package. More packages can provide the same virtual packages, so if one package has the editor as a dependency, one of the packages that provides editor is enough to install. These provisions are stored in the same file as the dependencies, along with other types of connections between packages, such as recommendations and suggestions [11].

In the software package network generated by cxnet, the packages are the vertices and these connections are the edges. It stores two types of packages (real, virtual) and three types of connections (dependencies, recommendation, provision). This is a directed network:

1 if A package *depends* on B package, there is a first type edge from A to B (A→B),

2 if A package *recommends* B package, there is a second type edge from A to B,

3 if A package *provides* the virtual package B, there is a third type edge from B to A.

The direction of the arrow in the third case has been chosen to be retrograde in the sense that in this case the connection information is listed at the package at the arrow head (at the target of the edge). The cxnet package uses colours on the plots to distinguish between the types of the vertices and edges, as can be seen in Fig. 1.

The software package network of the Debian distributions and its evolution are studied in the referenced articles [12, 13]. These analyses include only the real packages and the dependencies. The cxnet package can create and store a more detailed network of the connections between packages. It is a good exercise for the students to study how the properties of the network changes with the presence and absence of the virtual packages.

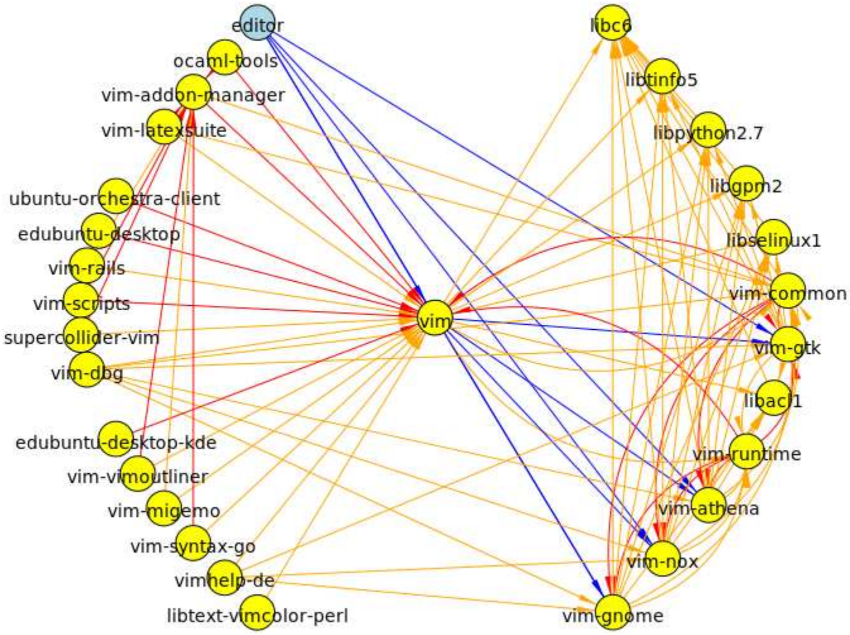


Figure 1

The neighbourhood of the package of the vim text editor

4 Investigating the Structure and the Evolution of Complex Networks

The structure of a diverse set of complex networks can be studied with cxnet. These networks include the software package network detailed earlier, the co-authorship networks of several fields of physics, the Internet at the autonomous system level, part of the Worldwide Web, the neural network of a worm, and the network of protein-protein interaction in yeast. However, cxnet is able to download network files from its webpage. The recent version of igraph (0.6) has its own network repository [14] as an alternative method to fetch networks to study. These networks include directed and undirected networks, bigger or smaller networks, and artificial and natural networks. Most of them are evolving networks, which means that new vertices and edges appear in this networks and some vertices can disappear as well, so the graph that describes the networks changes as the size of these networks grows. The evolution of complex networks can be studied using the thousands of software package networks of several versions of Ubuntu and Debian that have been stored.

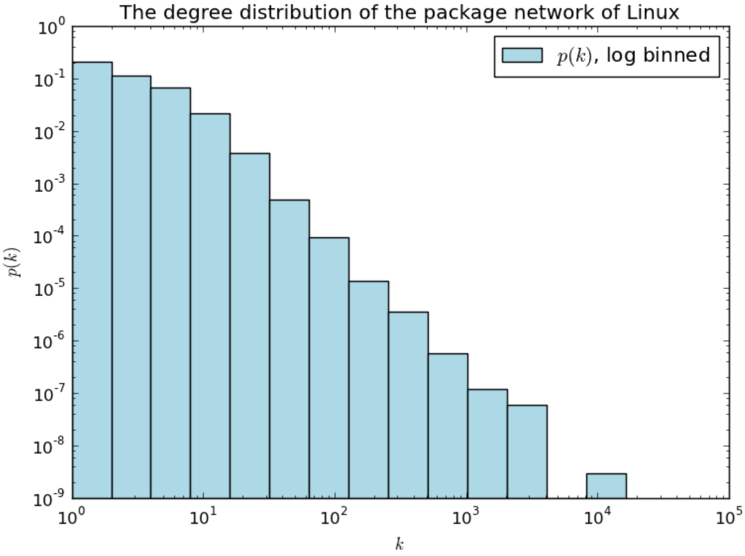


Figure 2

The degree distribution of the software package network. The degree is on the horizontal axis and the probability on the vertical one. Logaritmic binning (detailed in the text) has been used.

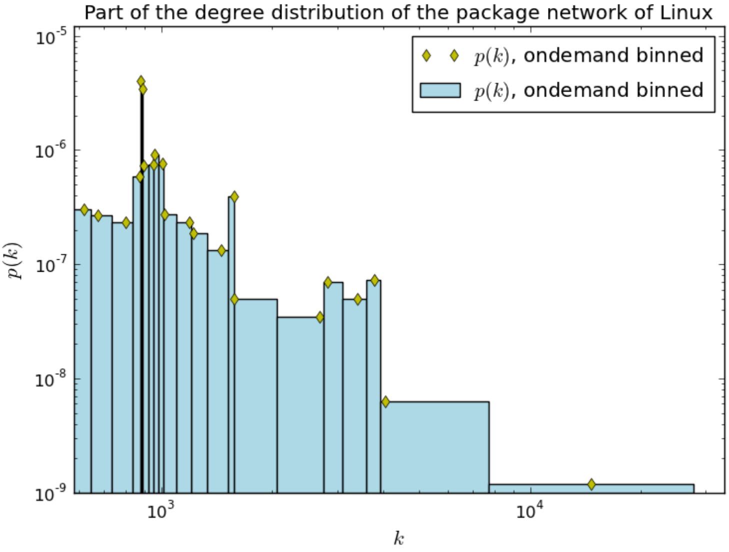


Figure 3

A part of the degree distribution of the software package network. Ondemand binning has been used. Detailed in the section 4.1.

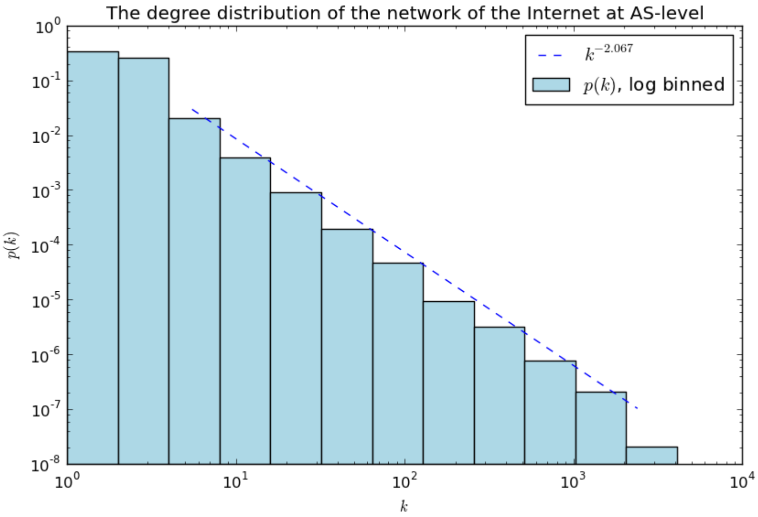


Figure 4

The degree distribution of the Internet. Logaritmic binning has been used.

4.1 Plotting the Degree Distribution

The degree distribution of a network is the p function that assigns to every natural number the probability that a randomly chosen vertex has this degree. The cumulative degree distributions of a network is the P function that assigns to every non-negative number the probability that a randomly chosen vertex has a degree equal or bigger than this value. In equations:

$$p(k) = Prob(\text{degree of the randomly chosen vertex} = k) \tag{1}$$

$$P(k) = Prob(\text{degree of the randomly chosen vertex} \geq k) \tag{2}$$

We can define indegree distribution or outdegree distribution and their cumulative analogously.

The plotting of the degree distribution (with or without the approximate power-law distribution) is easy with the cxnet package, as the examples in the appendix show. We can plot degree distribution or cumulative degree distribution. For the degree distribution we can choose three binning methods. In the first, the width of the bins is one degree. The second is logarithmic binning, which is widely used for plotting the degree distribution of the scale-free networks (Figure 2). The third is the method that is called the “ondemand” method by cxnet. In this method, the width of the bins depends on the degrees of the vertices. There is a division point between the degrees i and j if there is at least 1 vertex with degree i and at least 1 vertex with degree j, but there is no vertex with degrees between i and j. The division point is at the geometric mean of the two degrees ($\sqrt{ij}$). In Figure 3 a part of a degree distribution was plotted with ondemand binning using points and bars. The vertical edges of the bars are at the division points. The abscissas of the diamonds are the existing degrees in the network. In the cxnet program, we can set the binning method with the parameter values “all”, “log” and “ondemand”, respectively, as can be seen in the appendix. As we can see in Figure 4, the degree distribution of all networks (directed and undirected) can be plotted by the cxnet package.

4.2 The Clustering Coefficient Degree Diagram

The clustering coefficient is a number of the [0, 1] interval defined for the vertices of a network and for the whole network. For a vertex this coefficient is 1 if each of its neighbours is connected to the other neighbours and 0 if none of its neighbours are connected. The exact definition of the clustering coefficient of a vertex i is the quotient of the number E_i of existing edges between its neighbours and the maximal number of edges that could be between neighbours. In this definition, the network is handled as an undirected and simple one. The simple graph has no multiple edges between vertices and has no loop edge having the same vertex at its ends. The maximal edges between the k_i neighbours of vertex i is $(k_i \times (k_i - 1))/2$, so the clustering coefficient of the vertex i is

$$C_i = \frac{2E_i}{k_i \times (k_i - 1)} \tag{3}$$

The cxnet package is able to plot the clustering coefficient as a function of degree (Figures 5 and 6). The ordinate of the point in this plot is the arithmetic mean of the clustering coefficient of the vertices with the same degree. This function was investigated in the articles [15, 16] and it was concluded that the clustering coefficient in many real networks decreases roughly inversely proportional to the degree, and this is an indication of the hierarchical property of these networks.

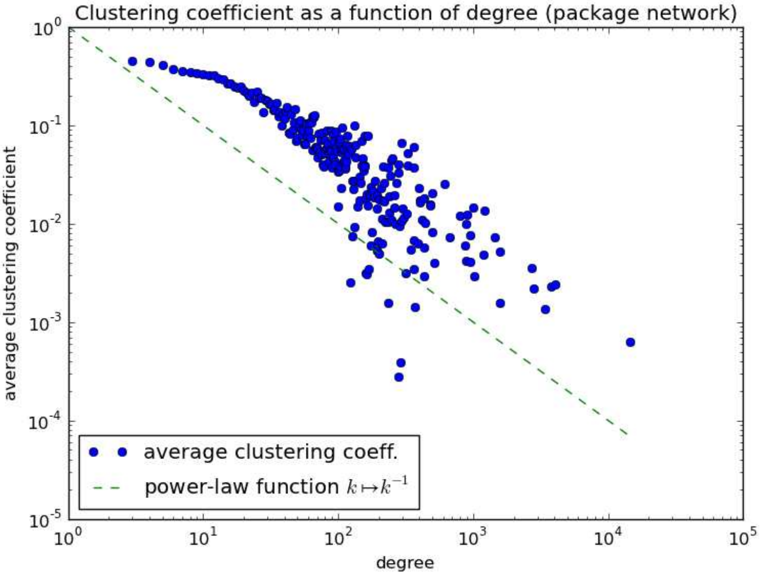


Figure 5

The average clustering coefficient of the software package network as the function of the degree

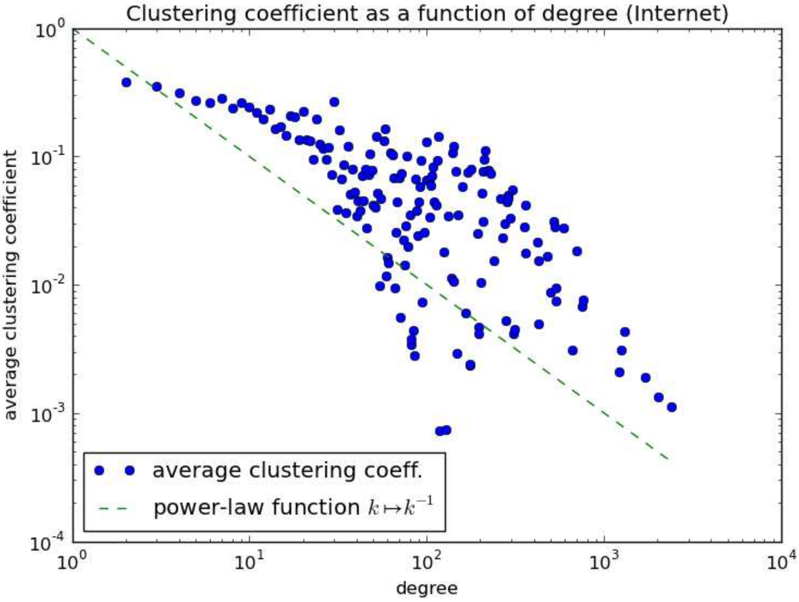


Figure 6

The average clustering coefficient of the Internet as the function of the degree

The clustering coefficient of the network is the arithmetic mean of the clustering coefficients of the vertices. In almost all networks the probability that there is an edge between two randomly chosen vertices is far less than the probability in the case that the two randomly chosen vertices would be chosen with the condition that the two vertices should have a common neighbour. In the example of a friendship network, the possibility that my two friends are friends of each other is more than the possibility that two members chosen from this network are friends. This property can be quantified with the clustering coefficient. The clustering coefficient of real networks is some order of magnitude higher than it would be if the edges would connect vertices randomly.

The clustering coefficient has no meaning for vertices with a degree of zero or one, so there are two possibilities to calculate the clustering coefficient of the network. In the igraph and cxnet packages, this coefficient is calculated as the average for the vertices with a degree more than one. Another possibility is also common in the literature: defining a value – usually 1 – as the clustering coefficient of these vertices, and the clustering coefficient of the network includes all vertices.

5 Documentation and Educational Materials

The cxnet package has a fairly up-to-date documentation on the webpage of the cxnet package [10]. This includes a detailed tutorial (English and Hungarian) and package reference for this package (English).

The educational materials for the “Complex networks” course, which was developed in our centre, are mainly in Hungarian. The syllabus and some video tutorials can be found on the web. Students can access tests on the e-learning (Moodle) website of Óbuda University. These tests cover the theory of complex networks, the software development in Python and the usage of the igraph, cxnet and pylab modules. Our earlier article [17] details some of the properties of the software package network, which can be studied with the cxnet module.

A book by Mark Newman [18] is big help for teachers and ambitious students. There are also some papers available for free on the web that include a summary of the most important concepts of complex networks as well as the properties of many real networks [1, 19, 20]. There are some other articles about more specific fields of complex networks that are also useful in education [15, 21–25].

Conclusions

The Python language with the cxnet package can be used to study networks easily, so it is an ideal tool for courses in higher education. On deb-package-based Linux distributions, the cxnet package is able to create the software package network of Linux, a large and fast evolving directed network. The most important network properties and the laws of network evolution can be studied with the cxnet packages using this and many other archived networks. This package is unique in the sense that it unifies some of the benefits of other packages without their drawbacks. These advantages are:

- the fast computation of the network properties for networks with millions of vertices as well,
- the plotting of the degree distribution with several binning methods or as a cumulative distribution,
- the plotting of the average clustering coefficient against the degree, to decide whether the network is hierarchical or not,
- the easy interactive investigation of the networks with the Python shell,
- the possibility to create the software package network of Linux.

As it is written in Python, the cxnet package gives the possibility for the students to implement algorithms easily in object oriented or procedural style.

Acknowledgement

I would thank to Tamás Nepusz for his help in using igraph. I would thank to Marianna Machata for her linguistic help and advice.

Appendix: Examples on the usage of the cxnet package

This is just a short introduction of the possibilities of cxnet. More detailed documentation can be found on the webpage of the cxnet package [10]. We need to start one of the python shells. We recommend to use the igraph with the `--pylab` option.

As the Network class of the cxnet is derived from the Graph class of igraph package, we can use all of its functions. The member functions (methods) of Network added in the cxnet starts with "cw", so we can distinguish between cxnet and igraph functions. We need to import the cxnet package, and then we can create or load the software package network:

```
  import cxnet
  net = cxnet.debnetwork()
```

or

```
  net = cxnet.load_netdata("ubuntu-12.04-packages-2012-07-18_21.06GMT.graphmlz")
```

Then we can count the number of vertices and edges:

```
  net.vcount(), net.ecount()
```

We can get the maximal degree and maximal indegree in the network:

```
  max(net.degree()), max(net.indegree())
```

We can plot the cumulative degree distribution of the package network (Figure 7).

```
  dd.cumulative_plot(with_powerlaw=True)
  legend()
```

We can create the degree distribution and plot with chosen binning methods and save the figure (Figure 8). As we want to create another figure we need to close the figure window, or close the figure:

```
  clf()                                                    # close the figure
  dd = net.cxdegdist()
  dd.set_binning("log")                                 # logarithmic binning
  dd.loglog()                             # plot using log scales on each axes
  dd.set_binning("ondemand")
  dd.loglog()
  savefig("degdist.png")                           # save the figure into a file
  dd.plot_powerlaw()         # power-law function with the calculated exponent
  print(dd.gamma)                  # print the absolute value of the exponent
```

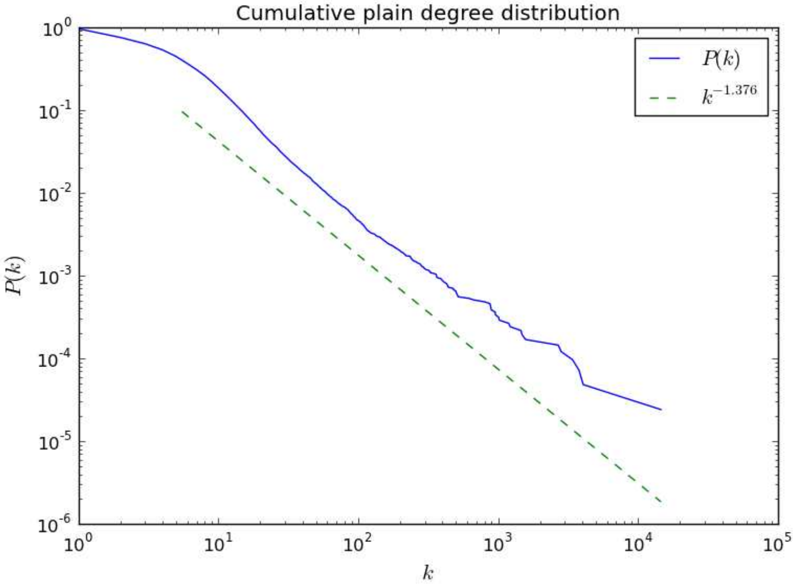


Figure 7

The cumulative degree distribution with the approximate power-law function

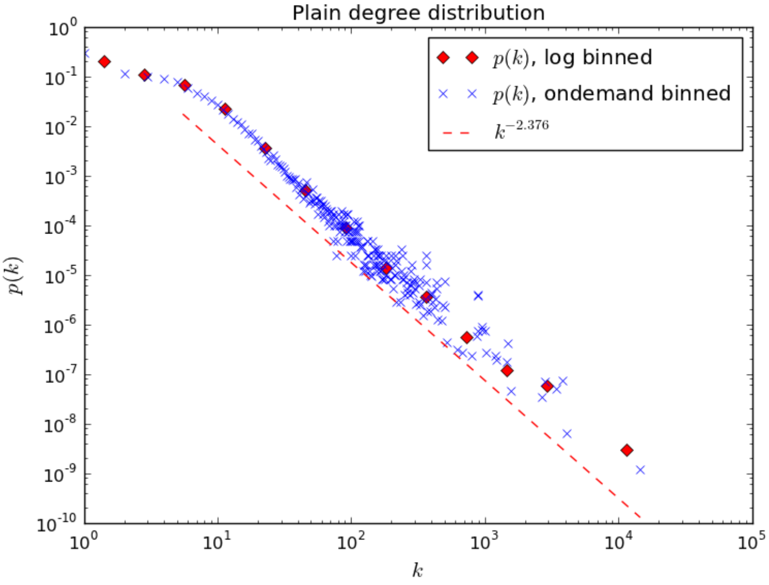


Figure 8

The degree distribution plotted using two binning methods and the approximate power-law function

If we typed dd = net.cxdegdist("in") in the previous example, the in-degree distribution would be plotted.

The bar plot allows to see the division points of the bins (Figures 2, 3 and 4). This plot can be plotted as follows:

```
clf()
dd.set_binning("log")
dd.bar_loglog()                           # logarithmic scales on both axes
legend()
```

We can list the packages with the largest degree (indegree, outdegree) as follows:

```
net.cxlargest_degree()
net.cxlargest_degree("in")
```

We can plot the average degree as a function of degree as follows (Figures 5 and 6):

```
net.cxclustering_degree_plot()
```

We can plot the neighbourhood of a package with one line (Figure 1):

```
net.cxneighborhood("vim")
```

The packages depending on the package called vim are at left, and the packages that vim depends on are at right. The vertices and edges are coloured according to its types. The blue vertex is virtual package, the yellow vertices are real packages. The gold edges are dependencies, the red ones recommendations and the blue ones provisions.

References

[1] Albert, R., Barabási, A.: Statistical Mechanics of Complex Networks, Reviews of Modern Physics, Vol. 74, No. 1, 2002, pp. 47-97

[2] Batagelj, V., Andrej M.: Pajek—Analysis and Visualization of Large Networks. *Graph Drawing*. Springer Berlin/Heidelberg, 2002

[3] Börner, K., et al. Rete-Netzwerk-Red: Analyzing and Visualizing Scholarly Networks using the Network Workbench Tool. *Scientometrics*, 2010, 83: 863-876

[4] Siek, J. G., Lee, L. Q., Lumsdaine, A.: *The Boost Graph Library: User Guide and Reference Manual*. Addison-Wesley Professional, 2001

[5] Mark Summerfield: Programming in Python 3: A Complete Introduction to the Python Language, Addison-Wesley Professional, 2010, see also http://python.org

[6] Csárdi, G., Nepusz, T.: The Igraph Software Package for Complex Network research, InterJournal Complex Systems, 2006, Manuscript Number. 1695

[7] Hagberg A. A., Schult D. A., Swart P. J.: Exploring Network Structure, Dynamics, and Function using NetworkX. , *Proceedings of the 7th Python*

in Science Conference (SciPy 2008), pp. 11-15, Pasadena, CA USA, August 2008

[8] Tosi, S.: Matplotlib for Python Developers, Packt Publishing, 2009, see also http://matplotlib.sourceforge.net

[9] The documentation of the python-apt program, http://apt.alioth.debian.org/python-apt-doc

[10] The homepage of the cxnet program, http://django.arek.uni-obuda.hu/cxnet

[11] Hertzog, R., Mas, R.: The Debian Administrator's Handbook, Freexian SARL, 2012, Chapter 5, see also, http://www.debian.org/doc/manuals/debian-faq/ch-pkg_basics.en.html

[12] Maillart, T, Sornette, D., Spaeth, S., von Krogh, G.: Empirical Tests of Zipf's Law Mechanism in Open Source Linux Distribution, Phys. Rev. Lett., Vol. 101, 2008, p. 218701, doi:10.1103/PhysRevLett.101.218701

[13] de Sousa, O. F., de Menezes, M. A., Penna, T. J. P. Analysis of the Package Dependency on Debian GNU/Linux. *Journal of Computational Interdisciplinary Sciences*, Vol. 1, 2009, pp. 127-133

[14] The Nexus homepage http://nexus.igraph.org

[15] Ravasz, E., Barabási, A.: Hierarchical Organization in Complex Networks, Physical Review E, Vol. 67, 2003, pp. 026112, doi:10.1103/PhysRevE.67.026112

[16] Vázquez, A., Pastor-Satorras, R., Vespignani, A.: Large-Scale Topological and Dynamical Properties of the Internet, Physical Review E, Vol. 65, No. 6, 2002, pp. 066130

[17] Horváth, A.: Studying Complex Networks with Cxnet, Acta Physica Debrecina, Vol. XLIV, 2010, pp. 37-47

[18] Newman, M. E. J: Networks, An Introduction, Oxford University Press, Oxford, 2010.

[19] M. E. J. Newman, The Structure and Function of Complex Networks, SIAM Review, Vol. 45, No. 2, 2003, pp. 167-256

[20] Dorogovtsev, S. N., Goltsev, A. V. and Mendes, J. F. F.: Critical Phenomena in Complex Networks, Rev. Mod. Phys., Vol. 80, No. 4, pp. 1275-1335

[21] Colizza, V., Barrat, A., Barthelemy, M., Vespignani, A.: Predictability and Epidemic Pathways in Global Outbreaks of Infectious Diseases: the SARS case study, BMC Medicine, Vol. 5, 2007

[22] Kitsak, M., Havlin, S., Paul, G., Riccaboni, M., Pammolli, F., Stanley, H. E.: Betweenness Centrality of Fractal and Nonfractal Scale-Free Model

Networks and Tests on Real Networks, Phys. Rev. E, Vol. 75, No. 5, Part 2, 2007

[23] Myers C. R.: Software Systems as Complex Networks: Structure, Function, and Evolvability of Software Collaboration Graphs, Phys. Rew. E, Vol. 68, 2003

[24] Zlatic V., Bozicevic, M., Stefancic, H., Domazet, M.: Wikipedias: Collaborative Web-based Encyclopedias as Complex Networks, Phys. Rew. E, Vol. 74, 2006

[25] Liu, Y.-Y., Slotine J.-J., Barabási, A.-L.: Controllability of Complex Networks, Nature, Vol. 473, No. 7346, 2011, pp. 167-173

Analysis of the Efficiency of Applied Virtual Simulation Models and Real Learning Systems in the Process of Education in Mechatronics

Slobodan Aleksandrov[1], Zoran Jovanović[2], Dragan Antić[2], Saša Nikolić[2], Staniša Perić[2], Radica Aleksandrov[1]

[1] Technical School Trstenik
Vuka Karadžića 11, 37240 Trstenik, Republic of Serbia

[2] University of Niš, Faculty of Electronic Engineering
Department of Control Systems
Aleksandra Medvedeva 14, 18000 Niš, Republic of Serbia
E-mail: zoran.jovanovic@elfak.ni.ac.rs, dragan.antic@elfak.ni.ac.rs, sasa.s.nikolic@elfak.ni.ac.rs, stanisa.peric@elfak.ni.ac.rs

Abstract: The rapid development of science and technology sets high demands for schools and faculties in terms of educating students to be able to manage complex technological systems. On the other hand, the application of new technologies in modern industry requires the creation of real and virtual laboratories capable of producing the conditions for the rapid and reliable transfer of new knowledge and skills from teachers to students. Having this in mind, the main focus of this paper is the realization of a module as a combination of virtual and real learning systems. The advantage of these systems is reflected in a possibility of creating virtual laboratories using three-dimensional models that simulate real industrial systems. Research shows that the use of virtual didactic systems in modern education increases motivation for learning, reduces learning time and enables modelling and simulation of real systems. However, in order to improve professional competencies, an interaction of simulation models with real industrial systems is needed.

Keywords: Mechatronics; education; learning systems; real and virtual laboratory; simulation

1 Introduction

Education in the field of mechatronics is a rather complex teaching process and requires the use of modern teaching methods and modern learning systems [1]. The term mechatronics is defined by Tetsura Mori: "*The word, mechatronics, is* *composed of 'mecha' from mechanism and the 'tronics' from electronics. That is to say technologies and developed products will now incorporate electronics more intimately and organically into mechanisms, making it impossible to tell where one ends and the other begins*" [2]. With the development of microprocessor technology and its application in the control of electro-mechanical systems, the original definition of mechatronics evolved, so Harashima, Tomizuka and Fukada, defined Mechatronics as: "*The synergistic integration of mechanical engineering, with electronics and intelligent computer control in design and manufacturing of industrial products and processes*" [2]. Despite efforts, there is still no definition that completely covers the meaning and application of mechatronics.

For successful education in mechatronics, modern learning systems are needed, ones which correlate with real industrial mechatronics systems [3]. These learning systems are characterized by modularization, flexibility, re-programmability and software support for modelling, simulation and networking. To enable students to be capable of applying acquired theoretical knowledge and practical skills to real industrial systems [4], these modern learning systems have to be realised with industrial components.

Many papers try to explore an alternate approach which employs simulation to the real industrial production equipment in learning control technology [5]-[8]. Software tools such as *MatLab*, *LabView*, *Cosimir Robotics*, *Cosimir PLC*, etc., are used to model and simulate the behaviour of real practice/didactic scenarios. A variety of learning environments implemented and tested in the publication "Mechatronics training in real and virtual environments" is one of the outputs from Marvel, a pilot project under the "Leonardo da Vinci" action programme for implementation of the European Community's vocational training policy. The main target groups for Marvel were students, trainees and instructors in vocational education and training [9], [10]. Simulation can be used to reduce the amount of time that students have to spend on executing real experiments [11], [12]. It is obvious that computers have to be used as tools to provide alternative source of learning material, but computer simulation cannot replace all forms of applied training [13], [14].

It is very common in teaching that the application of software solutions for simulation and modelling are potentiated, wherein the use of real learning systems is neglected. As a result of insufficient practice with real learning systems, after schooling, we get personnel that need additional training in working with industrial systems. On the other hand, the use of the real learning systems exclusively is limited by the small number of students that can be trained at the same time and a certain number of different learning systems and high education price. In order to make the teaching process challenging, stimulating and efficient, it is desirable to use modern software and hardware learning systems. In this way, the price is reduced, the number of participants in a group, trained simultaneously, is increased and different dynamics of knowledge and skills acquisition, according to the potential and desires of students, is enabled.

The new concept of education in this field should link theoretical knowledge and practical skills. In this way, these learning laboratories will enable students to design, control, test, and maintain existing and new mechatronics systems. A virtual environment created in this manner should enable students to respond to the future requirements of industrial mechatronics systems. On the other hand, these learning systems are characterized by rather high prices, so it is difficult for schools and faculties in our surroundings to provide sufficient number of different learning systems.

The goal of the research presented in this paper was aimed at the significance of the application of real and virtual learning systems in the teaching process of secondary vocational school of educational profile *Technician of mechatronics-pilot program*. The survey was carried out with the help of professors from the Faculty of Electronic Engineering in Niš. The aim of this cooperation is to establish the continual educational process, creating new teaching plans and programmes, based on the knowledge and skills that future students have acquired during previous schooling. The main advantage of this research is the fact that the two generations of students of the final year, from four different schools, have taken part in the survey. These four schools have exactly the same equipment, so the conditions for knowledge and skills acquisition are the same in all the schools. It must be highlighted that any other factors (social status of the student's family, the amount of time allotted to and needed for learning etc.) [1] which affect the teaching process and students are not taken into consideration in this step of the research. After the module was finished, based on their experiences in working with real and virtual systems, they evaluated the efficiency of the method of knowledge transfer.

This paper is organized as follows. In Section 2, we present the detailed description of the module. The short descriptions of real learning systems, used in the realization of this module, are presented in Section 3. In Section 4, a 3D simulation tool for practical PLC training, *Cosimir PLC*, is described. *Cosimir Robotics*, the leading 3D robot simulation system, is described in Section 5, and in Section 6, the survey is described and the obtained results are presented and discussed. In the last section, we give concluding remarks and future works.

2 Detailed Course Description

The curriculum of the module, *Testing and diagnostics of mechatronics systems,* is rather complex. In order to adopt the defined knowledge and skills covered by the teaching plan and program, previous knowledge of pneumatics, sensors, electric motors, microcontrollers and PLC is needed. The module is divided into three logical units. Sixty classes are assigned for the first part of the course.

In the introductory part, the students are divided into teams and acquainted with real mechatronics system in the mechatronics cabinet. The first mechatronics system is the *Festo MPS Distribution Station*. The students are familiarized with the procedure for linking a mechatronics system to a computer, installing the required software and transferring the program from a computer to PLC station for distribution. Next, the station is put into operation and the algorithm of functioning of the mechatronics system is monitored. After the demonstration, each of the teams repeats the procedure of demonstration. Every member of a team fulfills certain procedures on their own. There is only one distribution station in the mechatronics cabinet, so the other teams are familiarized with electric and pneumatic scheme, sensors and actuators, in the meantime. Five school classes are assigned for this exercise (E1) and for each of the following exercises.

The second exercise (E2) is an introduction to the software for 3D modelling and simulation, *Cosimir PLC*. This program allows for individual work of students, getting to know all components of the system, simulation of functioning of mechatronics system and monitoring the state of sensors and actuators. With the use of virtual models, it is possible for every student to acquire knowledge and skills on the computer on his own, with the desired dynamics.

In the third exercise (E3), the professor simulates various faults on the virtual model of *Distribution station*, and the students test virtual models, detect faults and remove the causes of faults. When a student completely masters the software packages for simulation, he is ready to safely use real the mechatronics system.

The fourth exercise (E4) involves working on a real system for distribution, the detection of electrical, mechanical and pneumatic components, installation and reinstallation, and conducting electric and pneumatic wiring. The fifth exercise (E5) involves the detection and removal of mechanical faults and faults on pneumatic installation and components. The sixth exercise (E6) involves the detection and removal of faults on electric and electronic components. Students prepare reports in electronic form for each of the exercises. Training for the mechatronics system *MPS Sorting station* is realised in the same way as training for *Distribution station* (E7-E12).

The second logical unit is training on the *MPS Robotic system*. For this part, 60 classes are assigned (E13-E24). During the first exercise, students are acquainted with a real industrial robot, the *Mitsubishi RV-2AJ,* and the control of robot is demonstrated (manual, computer control and automatic mode).

That is followed by a short course in programming a robot in programming language *Melfa Basic IV*, learning about a software for the 3D modelling and simulation of *Cosimir Robotics*, writing and testing programs in virtual environment and, at the end, the simulation and detection of fault in 3D environment (E14-E17).

In exercises E18-E22 students program and test the program in the programming environment *Robot Explorer* and transfer the program in robot controller and monitor functioning of industrial robot. In E23 and E24, students simulate faults on the industrial robot, and the faults are detected and removed.

In the third logical unit (35 classes) students perform synchronisation of functioning of *Robot station*, *Distributing station* and *Sorting station* (E25). They simulate and detect faults on the formed technological process (E26, E27), network mechatronics systems in LAN (E28), monitor system parameters and control mechatronics systems through LAN and WAN network (E29) and create SCADA system (E30, E31).

3 Description of Experimental Framework

As we mentioned before, the fast development of modern control systems for industrial processes requires modern learning systems, which, in their construction and characteristics include real industrial systems [15]-[17]. Modern learning systems make it possible for students to apply acquired theoretical knowledge on real systems, to get to know industrial pneumatic, electric and electronic components, hardware structures and programming of PLC, monitoring and adjusting of system parameters, and the principles of the functioning of complex mechatronics systems. Combining several such systems, a unique complex system can be created that fully simulates a real industrial process. Getting to know the characteristics and kinds of sensors, mechanical modules, programming industrial computers and controlling actuators on industrial systems presents the basis for education in the field of mechatronics.

In the mechatronics cabinet in the Technical school Trstenik, Serbia, we use learning systems which belong to the Festo modular production line MPS (Modular Production System) [18]. Many technologies are integrated to make these sophisticated production systems: mechanical, electrical, electronics and software. The main feature of these systems is that they represent real industrial systems that are realised with industrial components and are managed by industrial PLCs. The laboratory framework is equipped with *Robot station Mitsubishi Melfa RV-2AJ*, *Distribution station* and *Sorting station* (see Fig. 1) learning systems. These systems are supported by software for programming the PLC and robots, software for communication and software for 3D modelling simulation.

The first learning system, the *Mitsubishi Melfa RV-2AJ,* is a robotic hand with five degrees of freedom. Robot control is performed via the controller CR1-571. The controller has a modern 64-bit RISC/DSP processor (CPU), which allows simultaneous control of up to 6 axes. All sensors and actuators of the robot are connected to the controller. The robot controller via serial link RS232c is connected to a PC. The basic functions of the controller are: indirect interpolation, direct interpolation, 3D interpolation, palletization, subroutines, multi-tasking, conditional branching, speed control and optimal route connection. The joints of the robot are driven by AC servomotors, while the actuator is a pneumatic grip having photo sensors. Thanks to this learning system the students should be able to master hardware structure of the system, the interaction, control and programming of real industrial robots.

Learning systems for distribution and sorting, which are controlled by the PLC, are used to program, test and diagnose real mechatronic systems. These systems are controlled by Festo PLCs FC640 which have a network module and support for TCP/IP and which are directly connected to a local switch device by Ethernet cable. Each PLC has a unique IP address, thus enabling direct access through the computer network. The *Distribution station* separates work pieces from the storing place. The fill level of the storing place is checked by a one-way light barrier. A double-acting cylinder pushes work pieces out individually. The charger module grips the separated piece with vacuum gripper. Driven by a rotary drive, the arm of charger moves the work piece to transfer point on conveyor of the *Sorting station*. When diffuse sensor detects the work piece, conveyor is started and the stopper is activated. Sensors in the front of the stopper detect the work piece characteristics (black, red or metal). The work pieces are stored in the appropriate slides via sorting gates that are moved by short-stroke cylinders via diverting mechanism. A through-beam sensor monitors the filling level of the slides. The main advantage of these systems is reflected in their modularity; i.e. the funcional moduls, sensors and algorithms can be changed.

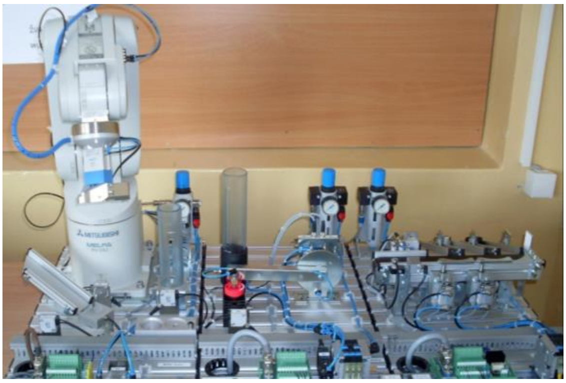


Figure 1
Experimental framework

It can be noticed that real learning systems are suitable for testing, diagnosing and servicing mechatronics systems because they enable hardware and software simulation of faults. However, the main shortcoming is that a maximum of three students can be trained on the same education system at the same time. Such systems for training have a relatively high price for most schools and faculties in the surroundings, so educational institutions are not able to own a larger number of the same educational systems on which simultaneous training could be held.

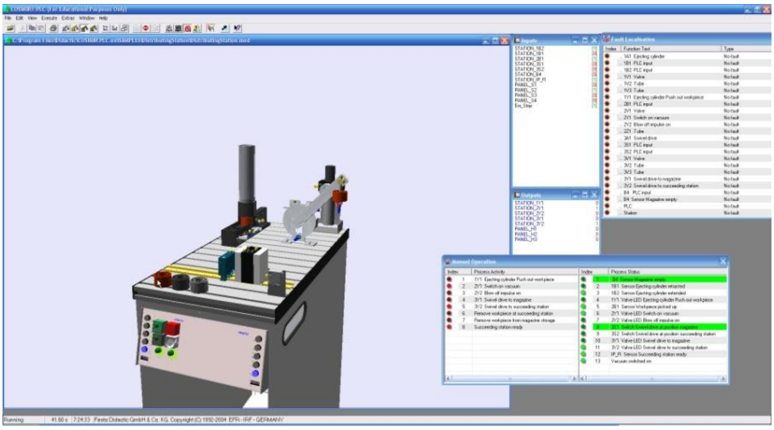
The alternative to the classical approach is to use virtual simulation learning systems, which simulate the functioning of real systems on a computer. Training is held on computers where programs for modelling and simulation are installed. In this way, a student can acquire the basic knowledge on their own, and, at the same time, groups from 10 to 20 members can be trained simultaneously.

4 *Cosimir PLC* Software

Nowadays, the use of the programmable logic device (PLD) and the programmable logic controller (PLC) is pervasive in educational and industrial applications [19]. It is already proved that the use of industrial units PLD and PLC can contribute to the student learning experience [20], [21]. Festo MPS mechatronics systems use the PLC of manufacturers known worldwide, like Festo, Siemens, Mitsubishi and Allen-Bradley, for controlling. In this paper, we present educational systems controlled by Festo PLC FC640. For realistic three-dimensional simulation of PLC system, the appropriate software *Cosimir PLC* is used.

Cosimir PLC represents a 3D simulation tool for practical PLC training. This virtual three-dimensional environment enables individual graphic work surroundings for each student. These software tools create an inspiring and stimulating work environment for the acquisition of new technologies. The basic characteristic of this software is that it possesses a library with finished models identical to the physical learning systems of the Festo MPS series. The procedure of starting, resetting and stopping the station fully corresponds to real systems. The software enables comfortable work, without the possibility of causing damage to the real system. *Cosimir PLC* does not require special hardware resources; therefore, all computers of the Pentium IV generation can be used. The newer version of this software is called the *Ciros Automation Suite*, but, for its optimal functioning, a computer of great processing power and with powerful graphics card and memory support is needed, which requires additional cost that is not small.

Virtual three-dimensional simulation represents real system with the only difference that control is performed by pressing a mouse taster on the model instead of the taster on the control panel. Simultaneously with the loading of three-dimensional model of mechatronics system, the corresponding program of the PLC that controls the functioning of chosen station is loaded. The work environment of the *Cosimir PLC* with the model of the distribution station is presented in Fig. 2.


COSIMIR PLC (For Educational Purposes Only)
File Edit View Execute Extras Window Help
C:\Program Files\Didactic\COSIMIR PLC ...\PLC15\DistributingStation\DistributingStation.mod
Inputs
Outputs
Fault Localization
Manual Operation

Inputs

STATION_1B2
STATION_1B1
STATION_2B1
STATION_3S1
STATION_3S2
STATION_B4
STATION_IP_FI
PANEL_S1
PANEL_S2
PANEL_S3
PANEL_S4
Em_Stop

Outputs

STATION_1Y1	0
STATION_2Y1	1
STATION_2Y2	0
STATION_3Y1	0
STATION_3Y2	1
PANEL_H1	0
PANEL_H2	0
PANEL_H3	0

Fault Localization

Index	Function Text	Type
●	1A1 Ejecting cylinder	No fault
●	1B1 PLC input	No fault
●	1B2 PLC input	No fault
●	1V1 Valve	No fault
●	1V2 Tube	No fault
●	1V3 Tube	No fault
●	1Y1 Ejecting cylinder Push out workpiece	No fault
●	2B1 PLC input	No fault
●	2V1 Valve	No fault
●	2Y1 Switch on vacuum	No fault
●	2Y2 Blow off impulse on	No fault
●	2Z1 Tube	No fault
●	3A1 Swivel drive	No fault
●	3S1 PLC input	No fault
●	3S2 PLC input	No fault
●	3V1 Valve	No fault
●	3V2 Tube	No fault
●	3V3 Tube	No fault
●	3Y1 Swivel drive to magazine	No fault
●	3Y2 Swivel drive to succeeding station	No fault
●	B4 PLC input	No fault
●	B4 Sensor Magazine empty	No fault
●	PLC	No fault
●	Station	No fault

Manual Operation

Index	Process Activity
● 1	1Y1 Ejecting cylinder Push out work piece
● 2	2Y1 Switch on vacuum
● 3	2Y2 Blow off impulse on
● 4	3Y1 Swivel drive to magazine
● 5	3Y2 Swivel drive to succeeding station
● 6	Remove workpiece at succeeding station
● 7	Remove workpiece from magazine storage
● 8	Succeeding station ready

Index	Process Status
● 1	B4 Sensor Magazine empty
● 2	1B1 Sensor Ejecting cylinder retracted
● 3	1B2 Sensor Ejecting cylinder extended
● 4	1Y1 Valve-LED Ejecting cylinder Push out workpiece
● 5	2B1 Sensor Workpiece picked up
● 6	2Y1 Valve-LED Switch on vacuum
● 7	2Y2 Valve-LED Blow off impulse on
● 8	3S1 Switch Swivel drive at position magazine
● 9	3S2 Switch Swivel drive at position succeeding station
● 10	3Y1 Valve-LED Swivel drive to magazine
● 11	3Y2 Valve-LED Swivel drive to succeeding station
● 12	IP_FI Sensor Succeeding station ready
● 13	Vacuum switched on

Running 41.60 s 7:24:33 Festo Didactic GmbH & Co. KG. Copyright (C) 1992-2004 EFR - IRF - GERMANY

Figure 2

Virtual three-dimensional learning system - MPS distribution station

There are also separate windows with statements of input and output of PLC on the desktop of the monitor. Starting the simulation, the values of input and output are changed according to the algorithm of functioning, and this change is shown in windows *Inputs* and *Outputs*. Monitoring of the sensors and actuators state, at any time, is performed by activating a special mode of this software, *Step by Step*. Monitoring and analysis of a system is possible only if the system is divided into modules, based on structure of control, mechanical configuration, electric motor drive, electric and electronic components, generator of signals and energy flow. In Table 1, the function modules that make up a structure of learning systems, together with their basic characteristics, are presented.

Table 1

Function modules and characteristics of real learning systems

No.	Function modules	Components, parts, programme, functions
1	Structure of system and control structure	Program flow chart, graphic diagram, functional diagram and system description
2	Mechanic configuration	Mounting and adjusting the mechanical parts, function mechanical modules
3	Drives	Electric, pneumatic, hydraulic, mechanical
4	Controlling elements	Electric, pneumatic, hydraulic, mechanical
5	Control system	Electrical relays, PLC, CNC, pneumatics, robot controllers
6	Generators of signals	Binary sensors, analogue sensors, digital sensors
7	Power supply	Electric, pneumatic, hydraulic

One of the most significant options of this software package is the possibility for simulation of a system fault. The process of fault simulation is performed in protected mode (*Teacher mode*), and to get to this function a password is needed.

A teacher chooses one or more faults from the range of possible system faults, such as the disconnection of the power supply, a sensor failure, a component missing, a fault in pneumatic installation, etc. Along with the choice of a fault, it is possible to define the duration of the fault as well as the time of its activation in the system. After the fault is set, we leave *Teacher mode* environment in order activate the fault. In order to test and diagnose the fault, there is a specific part of the program for fault detection – *Fault Localisation*. Using this command, we localise the fault, remove it and exit this part of program. The removed faults are marked with a green line, and the faults that are still active with a red line. All the system faults, with their type and time of appearance, are archived in a separate fault file (*Fault log*). These data can be reviewed and analysed after the simulation. After the fault has been removed, it is possible to start the system simulation.

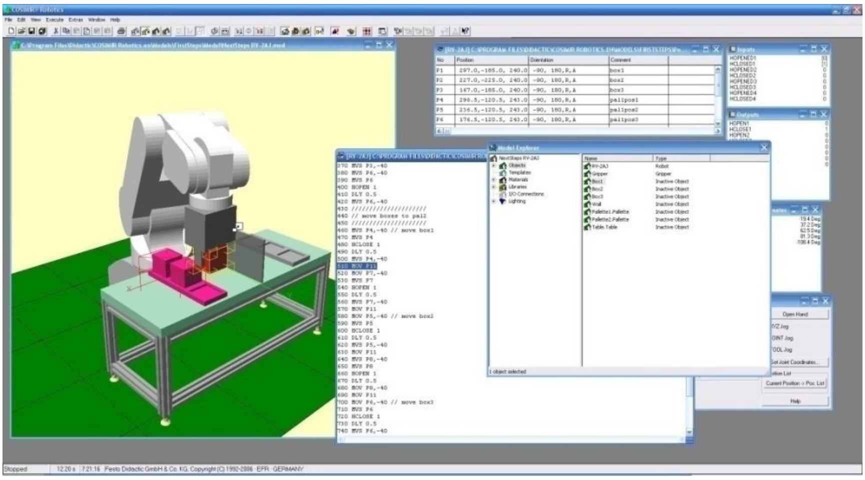
The procedure of fault simulation is similar on real learning systems. The addresses of input and output variables are the same, and the marking, appearance and position of system components are identical. Before real or virtual models are started, it is necessary to reset the system, i.e., to set initial conditions. To start a certain actuator, its respective sensors have to be active, as do all the states of all sensors relevant for that actuator change. This part of program enables monitoring process activity and states of all sensors and actuators. During the simulation of system, the statuses of input and output in PLC change, which is visually presented in the *Manual operation* window. There is a possibility to activate each of the actuators separately and in that way to monitor the status of signals which it activates or deactivates. In order to start this option, the simulation of system has to be closed. Thus, the system can be monitored every step of the way, which facilitates the learning process. This way of functioning corresponds to the manual mode of a real system, when a PLC program is executed step by step.

In this way, it is possible for students to become familiar with the basic components of mechatronics systems, monitor the algorithms of system functioning, program and test the PLC, simulate faults of a system, and detect and remove a fault. Using simulation software, students adopt much faster the principles of the functioning of a mechatronics system, as well as monitoring of system parameters and the statuses of input and output values on the mechatronics system.

5 *Cosimir Robotics* Software

A robot controller connected to a computer with serial link RS232 controls the robotic station. The robotic station is a part of the Festo MPS system, on which the industrial robot, *Mitsubishi RV2AJ,* with supporting work environment is set.

Due to specifics in controlling and programming the robot, the software development environment for programming and simulation, *Cosimir Robotics,* is used for this system. This software provides a virtual work environment for education in the field of robotics and automation. Attending the classes, the students have used the software for the simulation, testing and diagnostics of mechatronics systems in a virtual environment. Using this program, programs for the simulation of functioning of mechatronics systems can be created, and also, existing examples which are delivered with MPS systems can be used. We used the *Cosimir Robotics* toolbox for the programming and simulation of robot functioning. Mostly every school and faculty try to provide as many education systems (a minimum of three) as possible in order to create the conditions for learning a wide range of knowledge and skills. A virtual three-dimensional environment for real system simulation contains a library of models of real components, with the same labels, and it behaves as a real industrial system. The virtual system is intended for creating 3D models of robotic and flexible production systems, their programming and real operation simulation (see Fig. 3).



No	Position	Orientation	Comment
P1	297.0,-185.0, 240.0	-90, 180,R,A	box1
P2	227.0,-225.0, 240.0	-90, 180,R,A	box2
P3	167.0,-185.0, 240.0	-90, 180,R,A	box3
P4	296.5,-120.5, 243.0	-90, 180,R,A	palipos1
P5	236.5,-120.5, 243.0	-90, 180,R,A	palipos2
P6	176.5,-120.5, 243.0	-90, 180,R,A	palipos3

```
370 MVS P3,-40
380 MVS P6,-40
390 MVS P6
400 HOPEN 1
410 DLY 0.5
420 MVS P6,-40
430 //////////////////////
440 // move boxes to pal2
450 //////////////////////
460 MVS P4,-40 // move box1
470 MVS P4
480 HCLOSE 1
490 DLY 0.5
500 MVS P4,-40
510 MOV P11
520 MOV P7,-40
530 MVS P7
540 HOPEN 1
550 DLY 0.5
560 MVS P7,-40
570 MOV P11
580 MOV P5,-40 // move box2
590 MVS P5
600 HCLOSE 1
610 DLY 0.5
620 MVS P5,-40
630 MOV P11
640 MVS P8,-40
650 MVS P8
660 HOPEN 1
670 DLY 0.5
680 MOV P8,-40
690 MOV P11
700 MOV P6,-40 // move box3
710 MVS P6
720 HCLOSE 1
730 DLY 0.5
740 MVS P6,-40
```

NextSteps RV-2AJ
- Objects
- Templates
- Materials
- Libraries
- I/O Connections
- Lighting

Name	Type
RV-2AJ	Robot
Gripper	Gripper
Box1	Inactive Object
Box2	Inactive Object
Box3	Inactive Object
Wall	Inactive Object
Palette1.Palette	Inactive Object
Palette2.Palette	Inactive Object
Table.Table	Inactive Object

1 object selected

12.20 s 7:21:16 Festo Didactic GmbH & Co. KG, Copyright (C) 1992-2006 EFR GERMANY

Figure 3
Virtual robot environment – 3D model Mitsubishi RV2AJ

This facilitates planning, programming and testing of robots in every defined position, primarily in the 3D environment, and then in a real system. With this system, all sequences of moving, gripping and transferring are simulated, trajectory and time cycle is optimised, and elimination of collision is performed. The virtual environment is created by choosing an already-defined components library, such as machines, tools, warehouses, robots and other components.

The *Cosimir Robotics* software is based on the principle of an open approach to learning, which is characterised by constructive approach to learning; all models and tools are available and can be combined and used according to the individual's needs and the learning goals. This software is supported by multimedia system for providing help, *Robotics Assistant,* which contains textual description of robotic system, models, examples, a graphical presentation, as well as video material and animations. A very significant characteristic of this system is that three-dimensional virtual robot with its construction, components and functioning completely corresponds to a real industrial robot. The virtual environment enables defining the coordinate points through which the robot will pass, writing programs, compiling and simulation. Programs are tested in virtual the 3D environment, which enables students to monitor the program and motion of robot simultaneously. Thus, the effects of each command are visible and faults are easily detected. This system allows for different levels of knowledge and skills acquisition and promotes new methods of research, discovering and projecting. Using virtual models, students attain the basic knowledge and skills for working with real models. They come across problems when a designed system does not function. Faults can be detected independently, and it is possible to improve the efficiency of system functioning and to create more complex systems on the grounds of initial basic model. *Cosimir Robotics* enables monitoring tools coordinates, choosing the coordinate system, programming, testing the program functioning, and simulating the robot functioning in a 3D environment. Programming and controlling industrial robot RV2AJ is much easier, faster and safer, when students acquire the basic knowledge and skills using the simulation model.

It is very important to highlight that software packages for 3D simulation cannot replace real education systems, but they may facilitate and speed up the process of acquiring knowledge and skills, which will be confirmed in the conducted survey, presented in following section.

6 Survey Description and Analysis of Results

In order to verify the proposed teaching method, presented in this paper, an anonymous survey was conducted in four schools that have exactly the same conditions for realising the teaching process. The target group were the students of the first and second generation of the fourth year of technician of mechatronics. In the school year 2010/2011, 84 students (group A) were involved in the research, while in the school year 2011/2012, 102 students (group B) were involved. Each of the schools is equipped with the mentioned education systems and supporting software for virtual simulation in their mechatronics cabinets. This model of learning stimulates students to master very complex mechatronics systems in a modern, prompt and efficient way [9], [10], [22], [23].

In the course of implementing the new methods of learning and the new software and hardware tools, experts in the fields of pedagogy and psychology were consulted. The monitoring of the application of new methods and means of teaching is sporadic, so it often happens that new models are age-inappropriate, too complex, or inapplicable at all levels of learning. Since the student is at the centre of educational process, the way they accept new learning systems and the results they achieve in their application are very important.

The goal of this research is to get relevant information from the students about efficiency of virtual and real learning systems in mechatronics teaching. Based on these results, we can define certain contents which are adopted using virtual software for simulation, practical skills which must be adopted using real learning systems and record software models that do not give the desired results.

The students gave answers to the questions from Table 2 in the following way:

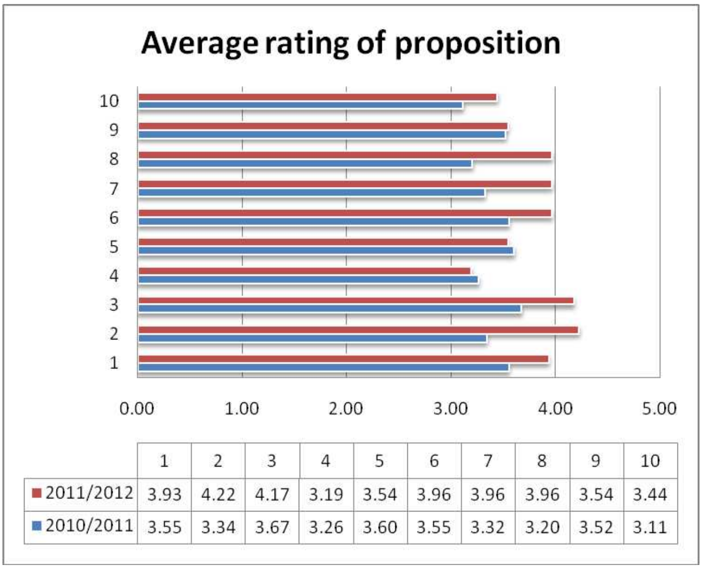
- Score 1 – Strongly disagree,
- Score 2 – Disagree,
- Score 3 – Neutral,
- Score 4 – Agree,
- Score 5 – Strongly agree.

Table 2

Survey statements used to assess learning performance

	Statement
1	The use of software for modelling and simulation is needed for successful knowledge and skills acquisition.
2	Software for simulation, *Cosimir PLC* and *Cosimir Robotics,* are interesting and challenging for students.
3	Software for simulation, *Cosimir PLC* and *Cosimir Robotics,* are suitable to show the principles of the functioning of real mechatronics systems.
4	Software for simulation, *Cosimir PLC* and *Cosimir Robotics,* are suitable for the simulation and detection of faults on mechatronics systems.
5	Software for simulation, *Cosimir PLC* and *Cosimir Robotics,* cannot fully replace real learning systems.
6	Software for simulation, *Cosimir PLC* and *Cosimir Robotics,* in combination with real learning systems are the best approach for the process of learning.
7	Software for simulation, *Cosimir PLC* and *Cosimir Robotics,* are easy to learn.
8	Training on real systems is easier if software for 3D modelling is used previously.
9	Software for simulation lessens the possibility of damaging real mechatronics systems.
10	Software for simulation reduces the student’s fear of complex real mechatronics systems.

Graphical representation of average score to any question from Table 2 is shown in Figure 4. These results give us the opportunity to analyse the application of education system in the school years 2010/2011 and 2011/2012, respectively.



	1	2	3	4	5	6	7	8	9	10
🟥 2011/2012	3.93	4.22	4.17	3.19	3.54	3.96	3.96	3.96	3.54	3.44
🟦 2010/2011	3.55	3.34	3.67	3.26	3.60	3.55	3.32	3.20	3.52	3.11

Figure 4

Graphic representation average rating of proposition in school years 2010/2011 and 2011/2012

To analyse the research, it is necessary to see also the results students have achieved at the end of their schooling. At their final examination, students took a test of vocational-theoretical knowledge and final practical work. The average score on the test of vocational-theoretical knowledge was 3.28, and the average score on the final practical work was 4.60. The average score on both practical and theoretical parts was 3.62. If we compare this score with average scores at all statements in the survey, we can see that deviation is very small. Experience in teaching practice has shown that students with better academic accomplishments have a wider range of interests and great knowledge in computers and programming and easily adopt and implement new methods and new means of teaching. Students with less knowledge mainly focus on work with real industrial systems. Percentagewise, the scores according to question is shown in Figs. 5 and 6.

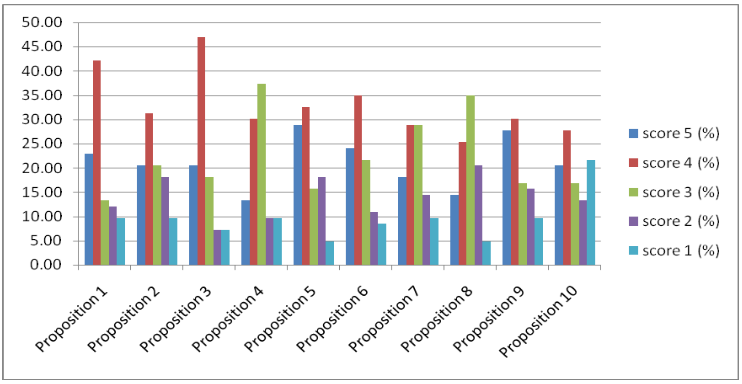


Figure 5
Percentagewise scores representation in 2010/2011

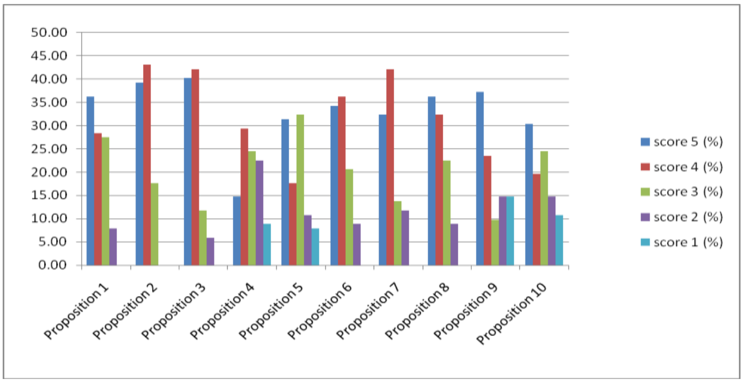


Figure 6
Percentagewise scores representation in 2011/2012

Comparing the results, for group A (2011) and group B (2012), from Figs. 5 and 6, we have the following analysis:

- The students point out that the use of simulation models in the teaching process is very necessary (65% for both groups), while 8% of students of group B do not accept the use of these models, significantly less in comparison to 22% of students of group A. This result suggests that the students of group B posess much more knowledge in the area of information and computer technology, and they willingly accept participation in this teaching method.
- In group B, 82% of students believe that the use of virtual models increases learning motivation. This high percentage, in comparison to the first group, can be justified by the fact that students are satisfied with the teaching method, and one of the reasons could be the possibility that the

students performed the assignments on their own. It should be noted that teachers have feedback results from the previous generation, and they now improved their teaching process.

- In addition, both groups of students believe that the use of virtual models in testing and diagnostics is necessary (44%). However, this rather small percentage suggests that students have not yet fully mastered the teaching material and they are not prepared to use the acquired knowledge on their own.
- Students of group B claim that virtual models are easier to learn (75%:47%) and therefore they have much more involvement in the learning process (69%:40%), which indicates that the second generation of students more easily accept the new software tools and that they are ready to use the new technologies for improving their knowledge.
- A high percentage (61%) of students of group A believe that virtual models cannot replace the real system in the learning process, while this percentage is significantly reduced in group B (49%), which indicates that the second generation of students recognizes the importance of combining virtual simulation models with real learning systems, which can be very helpful for solving practical problems in the future.
- The confirmation of the good teaching approach comes from the high percentage of students of both groups who believe that it is necessary to combine virtual and real learning systems. This percentage is 10% higher in group B, which shows that for the successful acquisition of knowledge and skills in the areas of technology, it is necessary to combine virtual and real didactic components and systems.

It must be highlighted that many other factors could affect the survey results, but they were not taken into consideration for this research. These factors are, for example, the various psychological orientations, general learning abilities (the ability of self-regulated learning), and the degree of comprehension (general intelligence, verbal skills), as well as several environmental factors, etc. [1].

Having in mind the results of the research, we can say that a large majority of students use virtual models in combination with real learning systems. The application of simulation models has a special significance in the hybrid model of learning, that is to say, in combination with training on real learning systems and industrial systems. The significance of application of software for simulation and modelling of mechatronics systems is in the fact that it:

- allows knowledge and skills acquisition in a modern and fun way,
- allows independent learning in preferred dynamics,
- allows a large number of repetitions of the same procedures,
- speeds up the process of learning,
- increasing motivation for learning,
- reduces the cost of education,

- reduces the fear of using real components and systems,
- prepares the students for working with real components and systems,
- allows working with a large number of different mechatronics systems.

Software for virtual learning, despite its great characteristics, does not represent replacement for real learning systems, but is a powerful software platform which facilitates and accelerates the process of knowledge and skill acquisition. The fusion of theoretical knowledge, practical skills and virtual 3D models represents a model for education in the field of mechatronics that has given great results in the education of technicians of mechatronics.

Conclusions and Future Works

In this paper, we have presented a hybrid learning model using real and virtual learning environment in the field of mechatronics. This learning method is applied in the teaching process of secondary vocational school of educational profile for the technicians of mechatronics pilot program in Trstenik, Serbia.

Simulations of processes can be used to familiarize students with all the components and units of modern mechatronics systems, in an interesting and modern way. Moreover, they have a significant role in modelling and testing systems, in cases where work on real systems is dangerous or very expensive. In this way, students can create and test desired models on their own and analyse the results of their work. The results of students' programming, testing and diagnostics are immediately visible, which gives great stimulation for further learning and exploring. There is no danger of student injury or damage to the system in a virtual environment, so various models of real systems can be simulated, tested and diagnosed.

However, to achieve the desired level of professional competencies, based on our experiences and the survey conducted in this paper, the best approach is to combine virtual models with real industrial systems. In that way, students are able to apply the acquired theoretical knowledge on real systems, which will decrease the fear of working with these kinds of real industrial systems in the future. Also, the possibility of injury during the work with real systems is reduced to a minimum, and the motivation for learning is raised to a higher level.

The main shortcoming of this learning method is that virtual and real learning systems are available for students at fixed intervals. In our future work, we will solve this problem by forming a system for remote access to the learning systems in mechatronics cabinet using video cameras, high-speed internet and web browsers. Apart from education, the system for remote access and control via the internet is also applicable to real industrial mechatronics systems.

Acknowledgement

This paper was realized as a part of the projects TR 35005 and III 43007, financed by the Ministry of Education and Science of the Republic of Serbia for the period 2011-2014.

References

[1] P. Tóth: Learning Strategies and Styles in Vocational Education, Acta Polytechnica Hungarica, Vol. 9, No. 3, pp. 195-216, 2012

[2] R. H. Bishop: The Mechatronics Handbook, The University of Texas at Austin, Austin, Texas, 2002

[3] S. Aleksandrov, R. Aleksandrov, P. Simić: Usage of Modern Didactic System in Education in Area of Mechatronic Enginering, Proceedings of the 6th International Symposium of Technology, Informatics and Education for Learning and Knowledge Society, Čačak, Serbia, 3-5 June, pp. 607-612, 2011

[4] S. Aleksandrov, S. Čajetinac, D. Šešlija: Didactic System Festo-MPS–Sorting Station and its Application in Education in the Field of Mechatronics, Proceedings of 10th International Conference Research and Development in Mechanical Industy - RaDMI 2010, Donji Milanovac, Serbia, 16-19 September, pp. 549-553, 2010

[5] K. Yeung, S. Chow: The Modular Production System (MPS): an Alternate Approach for Control Technology in Design and Technology, In: IDATER 98 Loughborough University, Hong Kong SAR

[6] R.-E. Precup, S. Preitl, M. B. Radac, E. M. Petriu, C. A. Dragos, J. K. Tar: Experiment-based Teaching in Advanced Control Engineering, IEEE Transactions on Education, Vol. 54, No. 3, pp. 345-355, 2011

[7] W. H. Liao, S. C. Wang, Y. H. Liu: Generalized Simulation Model for a Switched-Mode Power Supply Design Course Using MATLAB/ SIMULINK, IEEE Transactions on Education, Vol. 55, No. 1, pp. 36-47, 2012

[8] D. G. Lamar, P. F. Miaja, M. Arias, A. Rodriguez, M. Rodríguez, A. Vázquez, M. M. Hernando, J. Sebastián: Experiences in the Application of Project-based Learning in a Switching-Mode Power Supplies Course, IEEE Transactions on Education, Vol. 55, No. 1, pp. 69-77, 2012

[9] F. W. Bruns, H. H. Erbe, M. Faust: Engineering Future Laboratories, In: Marvel – A Leonardo da Vinci Pilot Project– Mechatronics Training in Real and Virtual Environments, Bremen, 2005

[10] D. Muller: Designing Learning Spaces for Mechatronics, In: Marvel – A Leonardo da Vinci Pilot Project– Mechatronics Training in Real and Virtual Environments, Bremen, 2005

[11] C. E. Pereira, S. Paladini, F. M. Schaf: Control and Automation Engineering Education: Combinig Physical, Remote and Virtual Labs, Proceedings of the 9th International Multi-Conference on Systems, Signals and Devices – SSD 2012, Chemnits, Germany, 20-23 March, pp. 1-10, 2012

[12] R. C. Hsu, W. C. Liu, Project-based Learning as a Pedagogical Tool for Embedded System Education, In Proceedings of 3rd International Conference on Information Technology: Research and Education, ITRE 2005, Hsinchu, Taiwan, 27-30 June, pp. 362-366, 2005

[13] I. G. Pop, V. Matiec: Transdisciplinary Approach of the Mechatronics in the Knowledge-based Society, Advances in Mechatronics, H. Martinez-Alfaro (Ed.), InTech, 2011

[14] L. Izsó, P. Tóth: Applying Web-Mining Methods for Analysis of Student Behaviour in VLE courses, Acta Polytechnica Hungarica, Vol. 5, No. 4, pp. 79-92, 2008

[15] E. Lindsay, M. Good: The Impact of Audiovisual Feedback on the Learning Outcomes of a Remote and Virtual Laboratory Class, IEEE Transactions on Education, Vol. 52, No. 4, 2009

[16] C. Buiu: Design and Evaluation of an Integrated Online Motion Control Training Package, IEEE Transactions on Education, Vol. 52, No. 3, pp. 385-393, 2009

[17] E. Montero, M. J. Gonzalez: Student Engagement in a Structured Problem-based Approach to Learning: A First-Year Electronic Engineering Study Module on Heat Transfer, IEEE Transactions on Education, Vol. 52, No. 2, pp. 214-221, 2009

[18] Festo, “The Modular Production System - User’s Manual”, 2013.
Available: http://www.festo-didactic.com

[19] I. Grout, J. Walsh: Microelectronic Circuit Test Engineering Laboratories with Programmable Logic, International Journal of Electrical Engineering Education, Vol. 41, No. 4, pp. 313-327, 2004

[20] J. J. Blakley, D. A. Irvine: Teaching Programmable Logic Controllers Using Multimedia-based Courseware, International Journal of Electrical Engineering Education, Vol. 37, No. 4, pp. 305-345, 2000

[21] L. Cheded, M. Al-Mulla: Control of a Four-Level Elevator System Using a Programmable Logic Controller, International Journal of Electrical Engineering Education, Vol. 39, No. 2, pp. 110-117, 2002

[22] M. Račić, J. Jovanović: Comparative Analysis of Mechanical and Mechatronic Design, Proceedings of the 9th International Conference Research and Development in Mechanical Industy - RaDMI 2009, Vrnjačka Banja, Serbia, 16-19 September, pp. 332-339, 2009

[23] S. H. Pulko, S. Parikh: Teaching Soft Skills to Engineers, International Journal of Electrical Engineering Education, Vol. 40, No. 4, pp 243-254, 2003

The Key Steps toward Automation of the Fixture Planning and Design

Attila Rétfalvi, Mihály Stampfer

Subotica Tech, Marka Oreškovića 16, 24000 Subotica, Serbia
ratosz@vts.su.ac.rs; stampfer@vts.su.ac.rs

Abstract: In automating fixture planning and design two major directions can be distinguished – automated dedicated fixture design, and automated modular fixture design. In this paper, the main steps that are needed for automated modular fixture planning and design are presented, and an integrated process planning and fixture design system – developed following these steps – is introduced. These main steps are feature recognition, systematization of the fixturing subtasks, defining the fixturing feature determination rules, systematization of the modular elements, and systematization of the element selection rules.

Keywords: modular fixture design; CAPP

1 Introduction

Automated fixture design means a process, in which, without human interaction, an appropriate fixture is designed for a given workpiece. Since nowadays the 3D modeling software is very widespread, as input for a system that is capable of automated fixture design should serve the 3D model of the workpiece, complemented with technological data like tolerances, surface roughness, etc. In the past there were numerous attempts to develop such a system, but full automation is not provided by any of the systems developed so far. More or less human interaction is needed in defining input data, and the user also has the surveillance role – he evaluates the solution. In this way, faults due to unforeseeable errors can be avoided.

Since different things are important for the designer, and different things are important for the process engineer, the design features can differ from manufacturing features; so if one wants to automate fixture design, a program is needed which can – on the 3D model of the workpiece – automatically find all machining and fixturing features and extract their relevant data. The next step should be a systematic evaluation of the extracted data and restructuring the data into a format that matches the needs of the fixture design. The entities of which the workpiece geometry is constituted should be examined according to fixturing aspects such as feasibility for supporting, for positioning and/or for clamping. When appropriate base surfaces are found, the system should look for appropriate fixturing elements, place them at the proper place and in this way build a feasible fixture for the given workpiece.

The system that is presented in this paper uses a neutral file format (IGES [1]) as input, in order to ensure platform independence, and recognizes the most common technological (manufacturing and fixturing) features. Then the user defines which surfaces should be machined, with what precision, and with what surface roughness; the user also prescribes the dimension, shape and relationship tolerances. The program examines the features and surfaces of the workpiece and tries to find the best supporting, positioning and clamping solution, taking in consideration the prescribed tolerances. Finally, it selects concrete modular fixture elements out from the database, and puts these elements on adequate place, building this way a fixture.

2 Literature Overview

Since during fixture design there are frequently recurring tasks and since the number of possible solutions is huge, process engineers have long been using computers to make their work faster and easier. There have been numerous attempts to develop programs that will automate, or at least considerably speed up, the fixture design. Some work has focused on automated feature recognition, others have focused on fixture construction, and there were also some who tried to solve both tasks.

One of the first systems, which builds a fixture from modular elements for machining a given workpiece was made by A. Márkus et al [2]. The supporting, locating and clamping bases were interactively given by the process engineer, and the system tried to build an appropriate fixture from some pre-assembled element combinations. The system ensures that the solution must not be more complicated than an available feasible more simple solution, it ensures that fixture elements are over the palette, and it checks the space usage and the interference. The user has to accept the proposed solution; if he refuses it, the system tries to find another solution. Chep et al [3] describe a method for restructuring the data stored in a CAD model into a form that is appropriate for CAPP systems for the automatic definition of machining operation sequence. These generalized data in the object-oriented database can also be used for the fixture planning and design. Prabhu et al [4] report a system that from 2D drawings (saved in DXF or IGES format) automatically extracts different kind of features, the dimensions and their attributes. The program has a text parser to interpret the different notes on the drawings. Arivazhagan et al [5] have developed a feature recognition system, which uses syntactic pattern recognition technique. Their machinable volume identifier program recognizes different kind of slots, blind slots and steps.

Subrahmanyam`s [6] heuristic-based volume decomposition method combines design feature usage (direct method) with a volume decomposition technique called heuristic slicing. In this way, the cell number (number of potential removal units) is halved compared with the cell-based approach, and the feature recognition time is decreased significantly. The fixturing algorithm in this research works only for single setup parts, and only flat surfaces are taken into consideration as potential fixturing surfaces. Kakish et al. in [7] outline a knowledge-based fixture design system. They focus on determining the design parameters and specifications of modular fixtures. Zhou et al [8] use the adjacent surface graph method in combination with feature tree reconstruction for feature recognition. This direct feature recognition system is able to find different kinds of slots, steps and holes. Rameshbabu et al [9] present a hybrid feature recognition method that combines volume subtraction and a face adjacency graph to cut recognition time. The feature identification is based on the number of feature faces and the adjacency count of each feature face. Their program recognizes different kinds of slots and steps. Wu et al [10] interpret the base principles of an algorithm that (with the assumption of having the primary positioning given) looks for the secondary and tertiary positioning points with the help of a linkage mechanism theory; the search for clamping points does a thorough investigation of IRC triangle and its same directed edges. The aim is to find a locating plan that ensures unambiguous, stable workpiece locating and where the loading and unloading is not hindered by the fixture elements. Boyle et al [11] studied the most recent works published about Computer Aided Fixture Design (CAFD), and they state that the CAFD is segmented in nature and that greater focus is needed on supporting detailed fixture design. They developed a case-based reasoning fixture design method. Xu et al [12] give an overview of the most commonly applied techniques in Computer Aided Process Planning, and they stress the need for integrated solving of the manufacturing related tasks and the importance of the environmentally conscious production. Alacron et al [13] developed a fixture planning and design system using the functional design theory. The user interactively prescribes the functional requirements, and the system, after defining the locating, supporting and clamping faces, selects modular elements and puts them in the defined place. Paris and Brissaud [14] present a process planning system that, after feature recognition assisted by an expert, associates machining processes to machining features and then organizes them into a global machining plan of the workpiece. Finally, it includes recommendations on fixturing features and determines the positioning quality, stability and cutter accessibility indices. Kumar et al [15] introduce an interactive fixture design system in which the user defines the fixturing surfaces and the program builds an interference free fixture. Perremans in [16] presents an expert system that in the possession of the fixturing features builds a modular fixture. In order to make the system manufacturer independent, he uses contact features, assembly features and tightening features to describe the modular elements. Vukelić et al [17] introduce an interactive combined (case-based and generative) system for drilling fixtures. Their system looks for an existing fixture solution in the database on the basis of the workpiece code; if does not find an appropriate existing solution, then it helps the user to generate a new fixture. Bansal et al [18] made an indirect feature recognition system that starts from the CAD model of the workpiece saved in STEP format. It determines the fixturing points by taking into consideration the prescribed tolerances, feature dependencies, manufacturing rules, fixture stability and ease of workpiece loading/unloading. In order to find the best fixturing points (which ensure the minimum tolerance deviances) the system slices the workpiece at different heights.

3 Feature Recognition

Feature recognition means the identification and extraction of the application relevant data from the part geometry [18]. Depending on the function, we distinguish design, manufacturing and fixturing features. The *design features* are groups of surfaces that are generated in a similar way; only some of their parameters differ – in this sense we can speak about protruded, swept, round, etc. features. *Manufacturing features* are those groups of surfaces that are made (usually removed) with the same tool or tool combination, with the same cutting parameters; they can also be those surface groups that serve as base surfaces during the assembly. In this sense machining features are different countersink and counterbore holes, slots, pockets, etc. *Fixturing features* are those surfaces or surface groups that are used for supporting, locating and/or clamping of the workpiece during the machining. The machining and fixturing features together are called *technological features*. As modeling software has evolved more and more commands for design feature generation have been built into them, but when one wants to use the features generated by them for process planning purposes, most often, he/she must first convert those into technological features. Since there are many different 3D modeling programs and since every program stores the data in a slightly different way, some neutral file formats (such as IGES, STEP, etc.) have been developed in order to ensure that a model made with one modeling program can be opened with other modeling programs, not only with the one with which it was made. When a model is stored in a neutral format, the design features are lost; only surface types (e.g. cylindrical, conical, etc.) are stored, so process planning – when it starts from a CAD model stored in neutral format – must begin with feature recognition.

The first step of feature recognition is the regeneration of the geometrical entities (points, curves, surfaces) that the model consists of and the classification and structuring of these data. The points are classified as start point, end point, midpoint, center point, point on curve, or point on surface. The curves consist of curve pieces, and these pieces are rated in three groups: straight line, conical sections and other, while curves are marked as inner or outer boundary curves.

The surfaces can be planar, cylindrical, conical and other. The orientation of the surfaces is also important information: the orientation can be horizontal, vertical or angled. The convexity property is important as well. Of course the characteristic data of each curve piece (such as length, radii, etc.) and surface (such as normal, size, etc.) are also determined. After this classification comes the identification of common curves and curve pieces of the surfaces. In the next step, the neighboring surfaces of the surfaces with common conical section are checked for type, and this continues until a terminating surface (a full conic or a non-cylindrical, non-conical surface that has not with actual surfaces coaxial cylindrical or conical surfaces) is reached. In this way surface groups are formed. Of course one has to leave out the found surface groups from further investigation in order that the same surface group not be found more times. Next, it is determined whether the found surface groups (potential features) form a feature and if yes, what kind of feature they form. Thus, we are looking for different kinds of holes (through or blind, with or without sinkage, with or without thread, with or without a slot for Saager-ring, etc.). Afterwards, different raised (boss) and sunken (pocket or slot) surfaces are looked for. Finally, it is determined if there are planar surfaces whose heights are the same, and could serve as supporting, locating or clamping surfaces. The feature recognition is introduced in more detail in [20].

4 Prescribing Technological Data

Technological data indicates those workpiece data that have influence on the kind of applicable machining process, on the tool dimensions and on the cutting parameters, and through these on the final fixture. Such data are, for example, the depth of a hole, the precision of the position of that hole, the diameter of the hole, the precision of the diameter, and the desired surface roughness. The geometrical data of each feature (such as the diameter, depth, position) are automatically extracted from the 3D model, but the precision of the position, the precision of the diameter, and the surface roughness are not stored when the model is stored in neutral format. So when all features are recognized, one must define which of them should be machined and with what precision; one must define the relationship tolerances too. One can take the recognized features one by one and prescribe the technological data. In the case of most features, the precision and the desired surface roughness must be prescribed. In the case of holes, in addition to the precision and surface roughness, one must define if it is threaded or not, and one must also define the initial state, i.e. whether it is pre-cast or drilled in full material. When there are more identical features (for example 5 counter sink holes with thread M6) the program will ask the user if all of them should be machined in the same way. Thus it is enough for the user to define her/his expectations one time: it is not necessary to do it five times. After that, the user selects the tolerance related features in pairs and prescribes the kind of relationship tolerance that binds the two features, as well as the precision of that tolerance. The prescribed precision of the relationship tolerances has a great impact on the structure and on the needed precision of the fixture.

5 Defining the Fixturing Type, Surfaces and Points

Since the system is developed for box-shaped parts, the first review is of the most common supporting, locating and clamping types used in the cases of box-shaped parts.

5.1 Types of Supporting, Locating and Clamping [19]

Box-shaped parts, especially gearbox casings, are most often machined on horizontal machining centers. Considering technological facilities of horizontal machining centers and analyzing existing clamping fixtures according to the position of the supporting surface of the workpiece, there are 3 types of supporting (Figure 1): (1) Horizontal (denoted with “pos1”), (2) Vertical (“pos2”), (3) Vertical with the ability of partial machining of the supporting face (“pos3”).

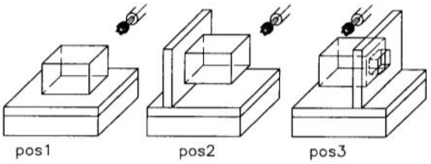
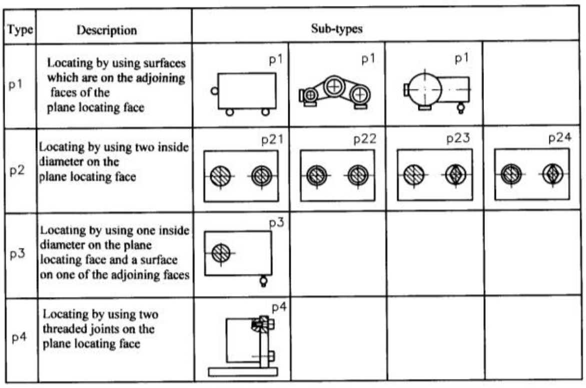


Figure 1
Supporting (Plane locating) types

There are 4 types of side locating (guiding) established (Figure 2): (1) side locating with the help of surfaces adjoining to the supporting face, (2) side locating with the use of two inside diameters on the supporting face, (3) side locating with the utilization of one inside diameter laying on the supporting face and one face adjacent to the supporting face, (4) side locating with the application of two threaded joints on the supporting face.

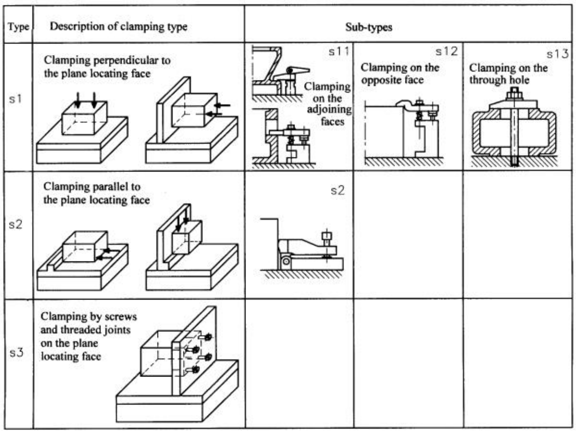
On the basis of the clamping force direction, one can distinguish between perpendicular clamping (s1), when the clamping force is perpendicular to the supporting surface, and parallel clamping (s2), when the clamping force is parallel with the supporting surface (Figure 3).



Type	Description	Sub-types			
p1	Locating by using surfaces which are on the adjoining faces of the plane locating face	p1	p1	p1	
p2	Locating by using two inside diameter on the plane locating face	p21	p22	p23	p24
p3	Locating by using one inside diameter on the plane locating face and a surface on one of the adjoining faces	p3			
p4	Locating by using two threaded joints on the plane locating face	p4			

Figure 2

Types of side locating



Type	Description of clamping type	Sub-types		
s1	Clamping perpendicular to the plane locating face	s11 Clamping on the adjoining faces	s12 Clamping on the opposite face	s13 Clamping on the through hole
s2	Clamping parallel to the plane locating face	s2		
s3	Clamping by screws and threaded joints on the plane locating face			

Figure 3

Types of clamping

The basic type s1, depending on the location of clamping faces, can be further divided into subtypes s11, s12 and s13. In the case of s11, the clamping surfaces are the closest parallel faces to the plane-locating (supporting) surface. In the case of s12, the clamping surface(s) is on the opposite side of the plane locating face. By s13 the clamping is carried out using a trough hole on the workpiece.

One special way of clamping is clamping by screws and joints on the plane locating face (s3). In this case, the clamping forces act perpendicular, but the force transmission happens in a different way.

The number of clamping points is also a very important characteristic of clamping. One distinguishes between clamping in one, two, three or four points. If the previous basic types are supplemented with this information, the results are the possible clamping types: s11_2, s11_3, s11_4; s12_2, s12_3, s12_4; s13_1, s13_2; s2_1, s2_2; s3_2, s3_3, s3_4. In the enumerated notation the last number denotes the number of clamping points.

5.2 Suitable Surfaces for Locating and Clamping

5.2.1 Suitability for Plane Locating

The suitability of a surface for plane locating depends on the shape and size of the surface. The best surfaces for plane locating are planar surfaces, then intermittent planar surfaces, surfaces on different parallel planes, cylindrical surfaces, and lastly a combination of cylindrical and planar surfaces.

The largest possible surface must be chosen for supporting the workpiece. The minimal allowed size of supporting surfaces is given as a percent of the largest dimensions of the part.

5.2.2 Suitability for Side Locating

a) *Suitability for side locating type p1*: Side locating can be divided into guiding and end supporting.

Suitability for guiding must be tested based on 3 aspects: shape, size and position of the surface.

- According to the shape of the guide locating element, the useable faces are: planar face, two planar faces, two cylindrical faces, a combination of cylindrical and planar surfaces, and single cylindrical surface.
- The typical dimension of a guiding surface must not be less than 35% of the longest dimension of the plane locating face.
- According to the position of the guiding element they must belong to adjoining faces of the plane locating face.

Suitability for end supporting must be tested based on two aspects: the shape and position of the surface.

- According to the shape of the surface, planar or cylindrical surfaces can be applied.
- According to the position, they must be on a face that is adjacent to both the locating and guiding faces.

b) *Suitable surfaces for side locating type p2*

According to the shape of the surface, there should be two holes on the plane locating face. The distance between the holes must not be less than 35% of the greatest length of the plane locating face.

Suitable surfaces for other side locating types are defined in a similar way.

5.2.3 Suitability for Clamping

The following principles must be respected during selection of clamping surfaces:

- The awaking forces should push the workpiece to the fixture.
- For the sake of rigid clamping, we must minimize the moments acting on the workpiece.
- The clamping should ensure positive rigidity.
- The clamping force must not deform the workpiece.
- The greatest shear force should be transmitted in a form-close way.

The clamping devices can be divided into two groups:

I. the clamping force changes with the deformation of the workpiece or the clamping device (screws, cams, wedges, springs, etc); or

II. the clamping force is constant (hydraulic, pneumatic, magnetic, etc.)

Of course the constant clamping force is better, but it can be achieved only with more expensive devices. So if the variation of clamping force is not too significant, one prefers elements from the first group.

Exactly which device will be used depends on the shape, size and location of the surfaces eligible for clamping.

For s1.1, the eligible surfaces are flat surfaces, through holes, or pockets; for s1.2 flat surfaces, intermittent planar faces or cylindrical surfaces; for s1.3 flat ring-like surfaces, flat frame-like surfaces or group of cylindrical surfaces (with axes perpendicular to the supporting face); for s2 flat surfaces, or intermittent planar faces; and for s3 threaded holes.

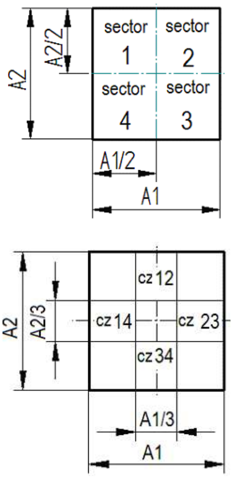
The location of clamping surfaces is on the opposite side to the supporting face at s1.1, s1.2 and s1.3; and at s1.3, the center of the hole is approximately coincident with the center of the supporting surface. At s2 there are flat faces opposite to the locating side, and they are near the supporting surface. At s3 there are threaded holes on the supporting surface and the distance between them is great enough (bigger than 40% of the height).

The size of the clamping face(s) should be large enough to ensure the settlement of clamping devices and to ensure appropriate clamping pressure. At s3, M6 or greater thread is needed.

In addition to these, one must check whether any of the above mentioned principles are violated.

5.3 Determination of the Clamping Points

For smaller parts, it is enough to clamp the part in two points, but a larger workpiece should be clamped in at least three points, and if there is enough space, it would even be advisable to do the clamping at four points. In order to distribute the clamping forces uniformly, the clamping area is divided into sectors and zones (Figure 4).



The position of the clamping points at clamping type **s11**, **s12** and **s3**	
Clamping in two points	• sectors 1,3 • sectors 2,4 • central zones 1-4, 2-3 • central zones 1-2, 3-4
Clamping in three points	• sectors 1,2 and central zones 3-4 • sectors 1,4 and central zones 2-3 • sectors 3,4 and central zones 1-2 • sectors 2,3 and central zones 1-4
Clamping in four points	• sectors 1,2,3,4 • central zones 1-2, 2-3, 3-4, 1-4

Figure 4

The division of the clamping area into sectors and zones

The layout of the clamping points for different cases can be seen in the table near the pictures. When determining the position of the clamping points, one must take care about the following: the clamping points must be over supporting points or as close to them as possible, and they must not be over features with precise tolerances.

6 Fixture Configuration

Fixture configuration means the selection of concrete elements for different fixturing tasks and the determination of their exact position and orientation.

The first step of fixture design is the systematization of the modular elements. Due to the high number of modular elements, the most commonly used elements for locating and clamping of box-shaped parts were chosen from the Kipp catalogue [21]. This reduces the search space and with it the searching time.

6.1 Systematization of the Modular Elements

The elements of a modular fixture can be divided into three groups:

a) base elements

b) functional elements

c) adapting elements

Base elements establish contact between the machine tool table and the fixture. This group includes different palettes, grid plates, and angle grid plates. (Table 1)

Functional elements are those elements that come in contact with the workpiece to fulfill a concrete function such as supporting, locating or clamping. In this group belong supports (Table 2), locators (Table 3), clamps (Table 5), and multifunctional elements (Table 4).

These groups are subdivided according to similarity of function into further subgroups, and a typical representative element of each subgroup is introduced in the tables. Every subgroup receives an ID in order to facilitate the referencing. These IDs are used in the next subsection (0), where the selection rules are systematically presented.

In order to reduce the number of needed elements, fixture manufacturers combine functions and thus offer multifunctional elements, too (Table 4). Of course, the smaller the number of the elements in a fixture, the more stable and precise it is.

Adapting elements (Table 6) are not always used. Sometimes the clamping surface is too high, or due to the shape of the workpiece, the supporting surface must be elevated or the locating elements lengthened. In such cases, one uses adapting elements to bridge the distance between the functional and the base elements.

The modular elements in the database are stored under coded names of the Kipp catalogue, where the first four plus three numbers refer to the type and subtype of the modular element. The following two numbers denote the size of the joining surfaces, while the last three or four numbers refer to some of the more relevant parameters of the element such as width, length or height. For example, in the code 8000 081 123240, the 8000 signifies that it is a grid plate with holes on every 50 mm. 081 indicates there are holes for socket head screws to facilitate fixing of the plate onto angle plate elements. 12 means the grid holes are M12 holes. 32 and 40 mean the width of the plate is 320 mm and length is 400 mm.

Table 1
Base plates

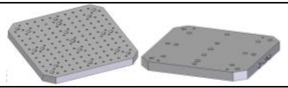
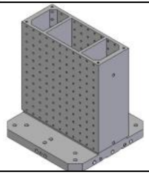
ID	Element	Name
EB1		Grid Plates
EB2		Angle grid plates

Table 2
Supporting elements

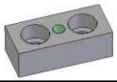
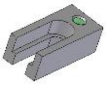
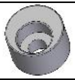
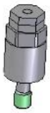
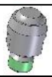
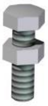
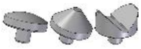
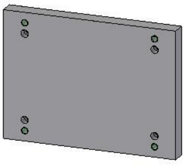
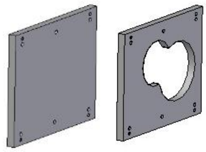
ID	Element	Name
ES1		Block supports
ES2		Adjustable supports
ES3		Eccentric supports
ES4		Jack Screw
ES5		Toggle Locators
ES6		Thrust Bolts
ES7		Jack Screw Tips
ES8		Adapting plate for locating bolts *(holes for bolts are machined according to the needs)*
ES9		Special plates

Table 3
Locator elements

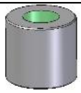
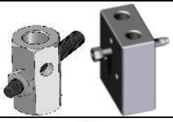
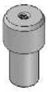
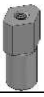
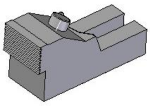
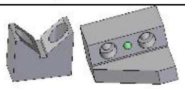
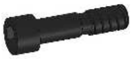
ID	Element	Name
EP1		Locating supports
EP2		Adjustable stops
EP3		Straight bolts
EP4		Flattened straight bolts
EP5		Side stops
EP6		V blocks
EP7		Locating bolts

Table 4
Multifunctional elements

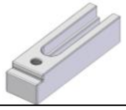
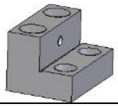
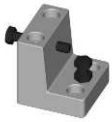
ID	Element	Name
EM1		Adjustable support with a step
EM2		Point Clamps
EM3		Adjustable Point Clamp

Table 5
Clamping elements

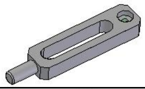
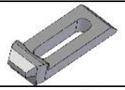
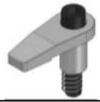
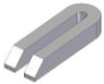
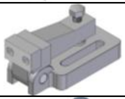
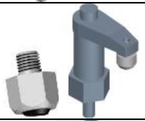
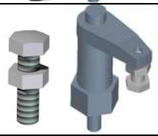
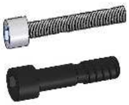
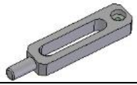
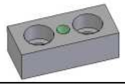
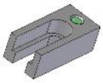
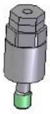
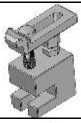
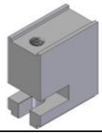
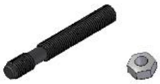
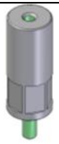
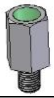
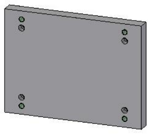
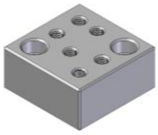
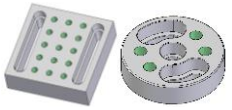
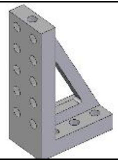
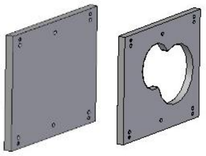
ID	Element	Name
EC1		Pin-End Strap
EC2		Goose-Neck Strap
EC3		Hook Clamps
EC4		Open Strap
EC5		Side Clamps
EC6		Toggle Locators
EC7		Thrust Bolts
EC8		Bolts

Table 6
Adapting elements

ID	Element	Name
EA1		Pin-End Strap
EA2		Block supports
EA3		Adjustabl e supports
EA4		Jack Screw
EA5		Clamp Holders
EA6		Extension Forms
EA7		Studs and Nuts
EA8		Joint Bars
EA9		Tie-Rod Bolts
EA10		Adapting plate for locating bolts *(holes for bolts are machined according to the needs)*
EA11		Height spacers
EA12		Adjustable spacers
EA13		Angle plates
EA14		Special plates

6.2 Systematic Presentation of the Selection Rules of Modular Elements

First the base element is selected in accordance with the size of the machine table and the type of the supporting (plane locating). Next the locators (and if needed, the adapting elements) are selected. Finally, the clamping (and if needed, adapting) elements are selected.

The selection rules are systematically ordered in matrices, and the columns contain the tasks. In the rows there are the elements and the conditions under which the elements can be used. **0** in the matrices means the element in question cannot be used for the task in question; **1** means that it can always be used; other symbols indicate under what circumstances it can be used. The nomenclature used is:

⌉ - NOT, V – OR, Λ – AND, PL - planar surface,

CYL - cylindrical surface, GPC - group of one planar and one cylindrical surface.

6.2.1 Selection of Base Elements

The selection of base elements depends on the dimensions of the machine tool's table, the dimensions of the workpiece, the type of supporting, the type of clamping and the type of locating. If the machine tool is defined, the base element's selection matrix looks as follows:

Table 7

Selection rules of base elements

		Supporting type		
Base elements		**pos1**	**pos2**	**pos3**
1	**EB1**	1	s3 V p4	1
2	**EB2**	0	⌉p4Λ⌉s3	0

The presented system first selects the type of the base element in the function of the proposed types of supporting, locating and clamping (Table 7) and then selects the appropriate size. Since the machine tool was established, only the workpiece size is taken into consideration. For example, if the (by SUPFIX module) proposed supporting type is pos2 (Fig. 2), the positioning type is p3 (Fig. 3), and the clamping type is s13 (Fig. 4), then the angle grid plate (EB2) is selected. However, if the supporting type is pos3, then a simple grid plate (EB1) together with angle plates (EA13) and special plates (EA14) should be used, in order to ensure tool approachability to those strictly connected surfaces of the workpiece, which lie on the supporting side of the workpiece. In the next step, the program verifies the dimensions of the workpiece and establishes the width and length of the base element. The main criterion is that at least one free row of grid holes must stay on each side after the workpiece is positioned on the center of the plate. This criterion ensures that there will be enough places for locating and clamping elements.

Grid plates and MC plates (EB1) can always be used for supporting type pos1 and pos3. For pos2, only when the clamping method is s3 or if the positioning type is p4. Angled grid plates (EB2) are used only for supporting type pos2, since they ensure greater rigidity, but can be used only when the positioning method is not p4 and clamping method is not s3.

The main elements of the code for selection of the base element are as follows:

```
base_element(ST,LT,CT,BE,) procedure (i,i,i,o).
selectSize(BET,SUT,W,L) procedure (i,o,o,o).
size1(wpD1,wpD2,SUT,W,L) procedure (i,i,o,o,o).
```

(ST – supporting type, LT - locating type, CT - clamping type, BE –base element, BET – base element type, SUT – sub type, W - width, L – length, wpD1 – the width of the workpiece measured parallel to supporting surface, wpD2 – the length of the workpiece measured parallel to the supporting surface)

```
base_element(“pos_2”,LT,CT,BE2):-   not(LT=”p4”),   not(CT=”s3”),   !,
selectSize(“EB2”,A,B,C), concat(“8000”,A,”12”,B,C,”.igs”,BE2).
  base_element(_,_,_,BE1):-                    selectSize(“EB1”,A,B,C),
concat(“8000”,A,”12”,B,C,”.igs”,BE1).

selectSize(“EB1”,A,B,C):- nx>ny, nx>nz, !, size1(yM,zM,A,B,C).
selectSize(“EB1”,A,B,C):- ny>nx, ny>nz, !, size1(xM,yM,A,B,C).
selectSize(“EB1”,A,B,C):- !, size1(xM,yM,A,B,C).
selectSize(_,A,B,C):- nx>ny, nx>nz, !,  size2(yM,zM,A,B,C).
selectSize(_,A,B,C):- ny>nx, ny>nz, !, size2(xM,zM,A,B,C).
selectSize(_,A,B,C):- size2(xM,yM,A,B,C).
```

The program first checks if the proposed supporting type is pos_2: if yes, then it verifies if the locating type is p4, and if not it verifies if the clamping type is s3; if not, then it selects the angled grid plate (EB2). In any other case, the grid plate (EB1) is selected. The concrete base element is selected in the function of the workpiece dimensions that are parallel with the supporting surface. For example, if the supporting surface’s normal vector points in x direction, then yMax and zMmax dimensions are considered. These data are obtained from the workpiece model during feature recognition (IPPO module) and are stored in appropriate variables.

6.2.2 Selection of Locating Elements

The locating task can be divided into two sub tasks, guiding and end stopping.

1.1.1.1 Selection of Guiding Elements

The selection of guiding elements first of all depends on the type of locating (Table 8). At types p2, p3 and p4 the subgroup for guiding elements is unambiguously determined, while at p1 the type of clamping and the type of the surface selected for guiding must also be taken into consideration.

For example, if one wants to solve p1 type locating, and the clamping type is not s2 and there is an eligible plane surface for locating, then it is possible to use two locating supports (EP1), or two adjustable stops (EP2), or one locating support and one adjustable stop. But if the only possible clamping method is s2, then some kind of side stops (EP5) should be used.

Table 8

Selection of guiding locators

Locators		Type of locating			
		p1	p2	p3	p4
1	**2 x EP1**	⌉s2∧PL	0	0	0
2	**2 x EP2**	⌉s2∧PL	0	0	0
3	**EP1 + EP2**	⌉s2∧PL	0	0	0
4	**EP3**	0	1	1	0
5	**EP5**	s2	0	0	0
6	**2 x EP5**	s2	0	0	0
7	**2 x EP6**	CYL	0	0	0
8	**EP6 + EP2**	GC	0	0	0
9	**EP7**	0	0	0	1

Table 9

Selection of end stop locators

Locators		Type of locating				
		p1	**p21** **p22**	**p23** **p24**	**p3**	**p4**
1	**EP1**	PL	0	0	0	0
2	**EP2**	PL	0	0	1	0
3	**EP3**	0	1	0	0	0
4	**EP4**	0	0	1	0	0
5	**EP5**	CYL	0	0	0	0
6	**EP6**	0	0	0	0	1

1.1.1.2 Selection of End Stops

The selection of end stops depends on the type of locating and also on the type of the selected end stop surface (Table 9).

6.2.3 Selection of Supporting Elements

The selection rules for supporting elements are shown in Table 10. It can be seen that in the majority of cases there is more than one acceptable solution. For example, if the supporting surface is a machined flat surface then different kinds of supports with flat supporting surfaces, or even grinded plates, can be used. What will be the determining factor is which of them fits to the workpiece dimensions best. When the supporting surface is larger, then block supports are preferred, while in the case of smaller supporting surfaces, adjustable or eccentric supports are the preferable choice. If the supporting surface is not machined yet, then toggle locators and thrust bolts or jack screws with tips must be used. So the system investigates the proposed type of supporting, locating and clamping, the number of proposed clamping points, and the size of the proposed supporting surface, and finally, it checks if the proposed supporting surface is machined.

6.2.4 Selection of Clamping Elements

Clamps are selected on the basis of the necessary type of clamping and on the basis of the type of the surfaces used for clamping (Table 11). The size of the clamping elements depends on clamping force needed and also on the vicinity of grid holes to the clamping place.

Table 10

Selection rules of supports

	Supports	Supporting surfaces				
		Machined plane surface	Group of machined plane surfaces at same height	Group of machined plane surfaces at two different heights	Raw plane surface, Group of raw surfaces	Group of raw plane surfaces at two different heights
1	**2 x ES1**	(p1 V p3) + (s2 V s4)	(p1 V p3) + (s2 V s4)	0	0	0
2	**4 x ES1**	(p1 V p3) + s4	(p1 V p3) + s4	0	0	0
3	**2 x ES2**	(p1 V p3) + (s2 V s4)	(p1 V p3) + (s2 V s4)	0	0	0
4	**4 x ES2**	(p1 V p3) + s4	(p1 V p3) + s4	0	0	0
5	**3 x ES3**	(p1 V p3) + s3	(p1 V p3) + s3	0	0	0
6	**4 x ES3**	(p1 V p3) + s4	(p1 V p3) + s4	0	0	0
7	**ES1+ES4**	0	0	(p1 V p3) + s2	0	0
8	**ES2+ES4**	0	0	(p1 V p3) + s2	0	0
9	**2 x ES3 +ES4**	0	0	(p1 V p3) + s3	0	0
10	**ES8**	p2 V p3	p2 V p3	0	0	0
11	**2 x ES8**	0	0	p2	0	0
12	**ES9**	p4 V pos3	p4 V pos3	0	0	0
13	**3 x ES5**	0	0	0	p1 V p3	0
14	**3 x ES6**	0	0	0	p1 V p3	0
15	**2 x ES5 +ES6**	0	0	0	0	p1 V p3
16	**2 x ES5 +ES7**	0	0	0	0	p1 V p3

* here the numbers beside the letter **s** denote the number of clamping points

The number of clamps is equal to the number of necessary clamping points that are prescribed (by SUPFIX) in the conceptual solution of the fixture. It may happen that in different points different types of clamps are needed. Toggle locators (EC6) and thrust bolts (EC7) are used to increase the clamping range of hook clamps (EC3) at s12 type clamping. The usage of hook clamps is restricted by the grid hole positions: that is, they can be used if there is a grid hole (G HOLE) close enough to the proposed clamping point. If one wants to clamp the workpiece on an inner cylindrical surface, a pin-end strap (EC1) is to be used.

6.2.5 Selection of Adapting Elements

Adapting elements are configured so that they can be connected not only to the base and functional elements, but also to each other. Table 12 shows how the adapting elements can be combined with each other. Table 13 shows which functional element and adapting element can be combined. Theoretically the number of adapting elements used in a chain that connects a base and a functional element can be infinite, but of course the goal is to reduce the number of adapting elements to as few as is possible. The selected functional elements determine the type of applicable adapting elements, and the main task is to find a dimension combination that ensures the minimal number of chain elements.

Table 11

Selection of clamps

Clamps		Type of clamping				
		s11	**s12**	**s13**	**s2**	**s3**
1	**EC1**	HOLE	1	0	0	0
2	**EC2**	0	1	0	0	0
3	**EC3**	G HOLE	G HOLE	0	0	0
4	**EC4**	0	0	1	0	0
5	**EC5**	0	0	0	1	0
6	**EC6**	0	1	0	0	0
7	**EC7**	0	1	0	0	0
8	**EC8**	0	0	0	0	1

Table 12

Mateability of adapting elements with other adapting elements and with base elements

Adop. Elem.	Adapting and Base Elements															
	A 1	**A 2**	**A 3**	**A 4**	**A 5**	**A 6**	**A 7**	**A 8**	**A 9**	**A 10**	**A 11**	**A 12**	**A 13**	**A 14**	**EB 1**	**EB 2**
A1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
A2	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	1
A3	0	1	0	0	0	0	0	0	0	0	1	1	0	0	1	1
A4	0	1	1	1	0	0	0	1	0	0	1	1	0	0	1	1
A5	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1
A6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1
A7	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1
A8	0	1	0	0	0	0	0	1	1	0	1	1	0	0	1	1
A9	1	1	1	1	0	0	0	0	1	0	1	1	0	0	1	1
A10	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1
A11	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1
A12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
A13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
A14	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0

Table 13

Mate ability of Functional elements with Base and Adapting elements

Func. Elem.	**Adapting and Base Elements**															
	A 1	**A 2**	**A 3**	**A 4**	**A 5**	**A 6**	**A 7**	**A 8**	**A 9**	**A 10**	**A 11**	**A 12**	**A 13**	**A 14**	**EB 1**	**EB 2**
ES1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1
ES2	0	1	0	0	0	0	0	0	0	0	1	1	0	0	1	1
ES3	0	1	1	0	0	0	0	1	0	0	1	1	0	0	1	1
ES4	0	1	1	0	0	0	0	1	0	0	1	1	0	0	1	1
ES5	0	1	1	0	0	0	0	1	0	0	1	1	0	0	1	1
ES6	0	1	1	0	0	0	0	1	0	0	1	1	0	0	1	1
ES7	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ES8	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1
ES9	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
EP1	0	1	1	0	0	0	0	1	0	0	1	1	0	0	1	1
EP2	0	1	1	0	0	0	0	1	0	0	1	1	0	0	1	1
EP3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
EP4	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
EP5	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1
EP6	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1
EP7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
EC1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
EC2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
EC3	0	1	0	0	0	0	0	1	0	0	1	1	0	0	1	1
EC4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
EC5	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1
EC6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EC7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EC8	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

When a functional element is selected, the system instantly checks if there is a need for an adapting element. From the characteristic dimension (usually the height) of the surface with which the functional element is in contact is subtracted the position of the surface of the ground element, in the direction of the surfaces normal. The difference is reduced with the distance between the contact surfaces toward the workpiece and toward the adapter element. This should be bridged with an adapting element or elements. Naturally, the range of the adjustability of the functional element is taken into consideration.

7 Test Example

In Figure 5 a gearbox casing and its most important dimensions and the prescribed tolerances are presented. Figure 6 shows the proposals of the SUPFIX module on the first setup, and the proposed solution is: lay the workpiece on the violet ring-like surface, locate it with the help of the pink inner cylindrical surface and the gray rectangular surface, and finally clamp it over the green ring-like surface.

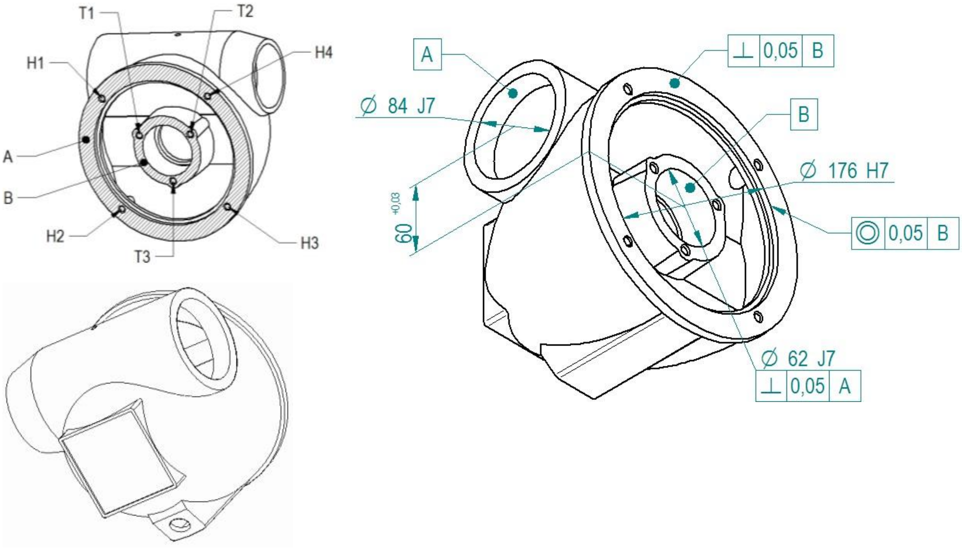


Figure 5

The workpiece, and the most important dimensions and tolerances

The proposed supporting type is “pos_1”, the side locating is “p3”, and the clamping type is “s_13”. In the first (auxiliary) setup, two ring-like surfaces (A,B) and four (H1,H2,H3,H4) plus three (T1,T2,T3) M10 threaded holes are machined. The (by FIXCO module) proposed fixture for this first setup and the Solid Edge assembly module (made by GLUE module) of the proposed fixture with and without the workpiece are also presented. The proposal for the second (main) setup in Figure 7 can be seen. During the second setup the workpiece should be laid on the violet ring-like surface, located with two threaded holes (green), and clamped over the four threaded holes (green). The proposed supporting is “pos_3”, locating is “p4”, and the clamping is “s3”. The main elements of the fixture without the workpiece and the whole fixture with the workpiece are shown.

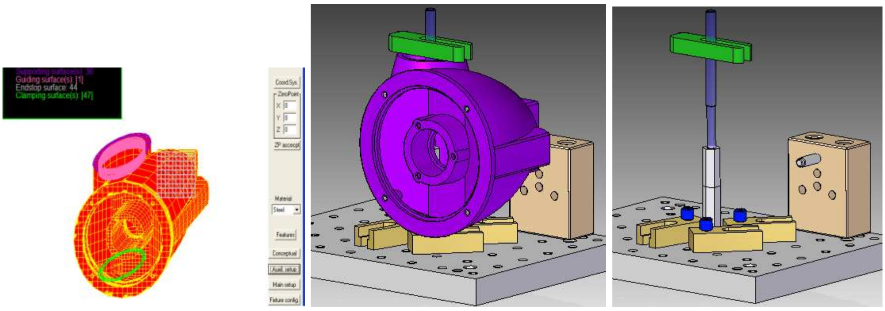

Supporting surface(s): 36
Guiding surface(s): [1]
Endstop surface: 44
Clamping surface(s): [47]

Figure 6

The proposed fixturing surfaces for the auxiliary setup, and the proposed fixture for the auxiliary setup

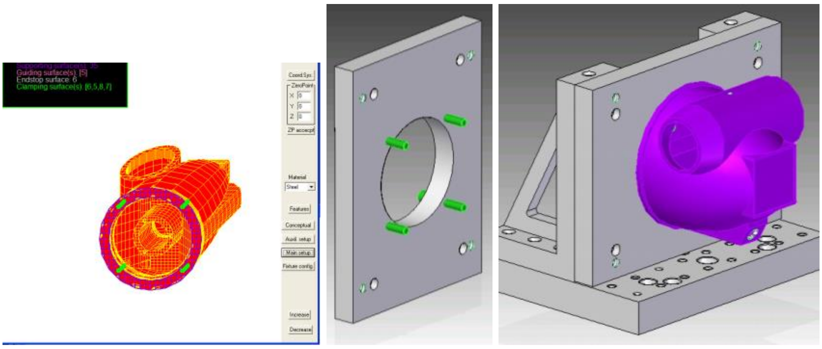

Supporting surface(s): 35
Guiding surface(s): [5]
Endstop surface: 6
Clamping surface(s): [6,5,8,7]

Figure 7

The proposed fixturing surfaces for the main setup, the adapter plate with clamping screws and the proposed fixture for the main setup

References

[1] US PRO (2008): Initial Graphics Exchange Specifications 5.3

[2] A. Márkus, Z. Márkusz, J. Farkas és J. Filemon (1984): Fixture Design using Prolog: An Expert System, Robotics and Computer Integrated Manufacturing 1 (2): 167-172

[3] A. Chep, L. Tricarico, P. Bourdet, L. Galantucci (1998): Design of Object-oriented Database for the Definition of Machining Operation Sequences of 3D Workpieces, Computers in Industrial Engineering 34 (2): 257-279

[4] B. S. Prabhu, S. Biswas, S. S. Pande (2001): Intelligent System for Extraction of Product Data from CADD Models, Computers in Industry 44:79-95

[5] A. Arivazhagan, N. K. Mehta, P. K. Jain (2009): A STEP AP 203 – 214-based Machinable Volume Identifier for Identifying the Finish-Cut Machinable Volumes from Rough Machined Parts, Int. J. Adv. Manuf. Technol. 42:850-872, doi 10.1007/s00170-008-1659-2

[6] S. R. Subrahmanyam (2002): A Method for Generation of Machining and Fixturing Features from Design Features, Computers in Industry 47: 269-287

[7] J. Kakish, P. L. Zhang, I. Zheid (2000): Towards the Design and Development of a Knowledge-based Universal Modular Jigs and Fixture Systems, J. of Intelligent Manufacturing 11:381-401

[8] X. Zhou, Y. Qiu, G. Hua, H. Wang, X. Ruan (2007): A Feasible Approach to the Integration of CAD and CAPP, Computer Aided Design 39:324-338

[9] V. Rameshbabu, M. S. Shunmugam (2009): Hybrid Feature Recognition Method for Setup Planning from STEP AP 203, Robotics and Compurter Integrated Manufacturing 25: 393-408

[10] Y. Wu, S. Gao, Z. Chen (2008): Automated Modular Fixture Planning Based on Linkage Mechanism Theory, Robotics and Computer Integrated Manufacturing 24: 38-49

[11] I. Boyle, Y. Rong, D. C. Brown (2011): A Review and Analysis of Current Computer-Aide Fixture Design Approaches, Robotics and Computer Integrated Manufacturing 27:1-12

[12] X. Xu, L. Wang, S. T. Newman (2011): Computer-aided Process Planning - A Critical Review of Recent Developments and Future Trends, International Journal of Computer Integrated Manufacturing 24 (1): 1-31

[13] R. H. Alacron, J. R. Chueco, J. M. P. Garcia, A. V. Idiope (2010): Fixture Knowledge Model Development and Implementation Based on a Functional Design Approach, Robotics and Computer Integrated Manufacturing 26:56-66

[14] H. Paris, D. Brissaud (2005): Process Planning Strategy Based on Fixturing Indicator Evaluation, Int J Adv Manuf Technol 25: 913-922

[15] A. S. Kumar, J. Y. H. Fuh, T. S. Kow (2000): An Automated Design and Assembly of Interference Free Modular Fixture Setup, Computer-aided Design 32:583-596

[16] P. Perremans (1996): Feature-based Description of Modular Fixturing Elements: the Key to an Expert System for the Automatic Design of the Physical Fixture, Advances in Engineering Software 25: 19-27

[17] Đ. Vukelić, U. Zuperl, J. Hodolic (2009): Complex System for Fixture Selection, Modification and Design, Int J Adv Manuf Technol 45:731-748

[18] S. Bansal, S. Nagarajan, N. V. Reddy (2008): An Integrated Fixture Planning System for Minimum Tolerances, Int J Adv Manuf Technol 38:501-513

[19] M. Stampfer (2009): Automated Setup and Fixture Planning System for Box-shaped Parts, Int J Adv Manuf Technol 45:540-552

[20] Rétfalvi A. (2011): IGES-based CAD Model Post Processing Module of a Setup and Fixture Planning System for Box-shaped Parts, IEEE 9th International Symposium on Intelligent Systems and Informatics, September 8-10, 2011, Subotica, Serbia

[21] Heinrich Kipp Werk: Workholding Systems, Catalogue of the Modular Fixture Elements

About Passwords

András Keszthelyi

Óbuda University, Károly Keleti Faculty of Economics, Institute of Organisation and Management, Népszínház u. 8, H-1081 Budapest, Hungary
E-mail: keszthelyi.andras@kgk.uni-obuda.hu

Abstract: In our age of cyber war and cyber crime, it is critically important to select and use "good" passwords to protect user accounts. A well-known general rule says that passwords should contain a mix of letters, numbers, and special characters. In this paper I will show mathematically that this rule is a misbelief. Instead of this, the length is the significantly important attribute. Then I will analyse the most common password structures and give an estimation on the time requirements of brute force attacks. (Un)fortunately there are a lot of password lists originating from a lot of intrusions and data thefts to analyse, and we have the incredible results of the latest brute force experiments. On the basis of these calculations we can state that passwords can give us strong protection if we apply some simple rules, unless the password encoding algorithm of the operating system is too weak. It is worth the time and energy for mathematicians to develop stronger hash functions and OS manufacturers to apply them, but this is not discussed here, and nor is how password using habits have changed.

Keywords: password; user authentication; password cracking; brute force attack; good password

1 Introduction

Latest news: Twitter was cracked at the beginning of February, 2013. In this data breach, the attacker(s) got access to the shadow passwords of about 250,000 users. Bob Lord, director of Information Security at Twitter said, "Make sure you use a strong password – at least 10 (but more is better) characters and a mixture of upper- and lowercase letters, numbers, and symbols... For more information about making your Twitter and other Internet accounts more secure, read our Help Center documentation or the FTC's guide on passwords." [1] At this point we have a really good occasion to think over what we know, sometimes incorrerctly, about the use of passwords and the attributes of a really good password.

It is a fact that nobody can reach a 100% security level in any field of life, especially not in IT. In our digital age, data security is very important (Google returns 13 million hints) in general, and in user authentication in particular.

“Passwords are a very poor authentication method. It is widely estimated that the majority of security breaches – as much as 80 percent – are attributable to persons picking “weak” passwords that are easy to guess or to stolen passwords that are compromised because of poor password protection practices. The method survives because it is still generally cheaper than the alternatives.” [2] As we shall see, passwords are *not* a “very poor” authentication method if used correctly. At least not poorer than a lost or stolen cellphone, token, etc.

The number of password-protected accounts an average user has is bigger than one would think at first glance, and it is increasing. According to Symantec, 44% of users have more than 20 password-protected accounts. [3] “The average user has 6.5 passwords, each of which is shared across 3.9 different sites. Each user has about 25 accounts (...) and types an average of 8 passwords per day.” [36]

As passwords will not disappear in the future, it is worth the time and energy to learn how to use them correctly and efficiently and to understand why some well-known rules and words of advice are misapprehensions, not to say “urban legends”, such as a “good” password looks like W@fK41#a&2s?.

2 Background

2.1 How to Get Somebody's Password

To store passwords in plaintext form is a serious security flaw. If anyone, anyhow can get access to the file plaintext passwords are stored in s/he will be able to personalize any of the regular users with the utmost ease. To avoid that, in most cases and in most systems the hash of the plaintext passwords are stored instead of the plaintext ones. Hash functions are one-way functions, which means that it is easy to calculate the hash of a plaintext password but it is impossible to calculate the plaintext password from the hash. These hashes are called shadow passwords. “Because the stored passwords cannot be deciphered, they are completely safe, even if the entire password file is (accidentally or maliciously) disclosed.” Denning said in 1982. [4] In those times perhaps there were no computational capacities which would have been enough to do brute force attacks against shadow passwords.

There are two different kinds of possibilities to get other users' password, stealing and guessing. *To steal* a password one can use any tools which are adequate to the circumstances. From hidden cameras to hardware keyloggers and trojan horse programs, there are a lot of possibilities, of which social hacking is not the worst: “Ninety per cent of office workers at London's Waterloo Station gave away their computer password for a cheap pen, compared with 65 per cent last year.” [6].

The other possibility is *to guess* the password. The simplest cases are when the password is a “default” one (e.g. password, asdfgh) or is in close connection with the login name (login – login12) or with the user's person (date of birth). A Frenchman even succeeded in breaking into Barack Obama's twitter account in 2010 (and there are other examples as well): “Cousteix managed to break into the accounts by searching information that is most commonly used for passwords, such as birth dates or pet names, on social networking sites. He lives with his parents and has no college degree, and has not had any special computer training.” [7]

“In 2008, the then-unemployed man was using Skype (...) when he dialed a random number and then entered the code "123456" (…) Although he didn't realize what he had done, the man was granted access to the French central bank's debt service.” [10]

2.2 Some of the Latest Bigger Password Thefts

2009, Hotmail. The list of stolen passwords initially contained 10,028 entries. After cleaning up the list, 9,843 valid passwords remained, of which 8,931 (90%) were unique. The most common password was: 123456. [13]

2009, Rockyou. In December, 2009 32 million passwords were revealed by a successful SQL injection attack. The passwords were stored in cleartext in the database, which is a serious case of carelessness. “The data provides a unique glimpse into the way that users select passwords and an opportunity to evaluate the true strength of these as a security mechanism. In the past, password studies have focused mostly on surveys. Never before has there been such a high volume of real-world passwords to examine.” [14]

“About 30% of users chose passwords whose length is equal or below six characters. Moreover, almost 60% of users chose their passwords from a limited set of alpha-numeric characters. Nearly 50% of users used names, slang words, dictionary words or trivial passwords (consecutive digits, adjacent keyboard keys, and so on). The most common password among Rockyou.com account owners is “123456”. [14]

2011, China. At the end of the year, about 100 million plaintext passwords were revealed from different Chinese websites. The passwords were stored as plain text in this case, too. Some interesting results of analysing the data are: “(1) users might choose less secure passwords for their convenience and ease of memorization, though their primary concern is online security; (2) for the same reasons, password reuse is common, as users tend to use the same passwords for multiple online accounts; and (3) passwords usually contain common words, or personal information, such as birthdays and family member names.” [15]

2012, Yahoo. In the summer a list of 450,000 usernames and plaintext passwords were revealed. Once again: plaintext passwords. According to [16], the top 10 passwords were: 123456, password, welcome, ninja, abc123, 123456789, 12345678, sunshine, princess, qwerty. The top 10 base words were: password, welcome, qwerty, monkey, jesus, love, money, freedom, ninja, writer.

2012, Philips. Only 400 real life shadow passwords were stolen. A researcher tried to crack these with an interesting result. He used John the Ripper to crack the passwords. The number of successfully found passwords shot up, then stabilized, and then remained steady. The first 25% of the passwords fell in 3 seconds; first half of them were found in 50 minutes; and only 53% in total after two hours. [37]

2012, LinkedIn. The data of 6,5 million users were stolen.

2.3 A Functional Approach to Password Usage

There are theoretical approaches based on entropy, but now let us prefer the practical point of view. We can state as axioms that a password *must* be not only hard to guess (for intruders) but easy to remember (for users). At first glance one would think that these two requirements are opposite to each other. A survey in 2010 points out the facm that users know this clearly, at least in theory, see Table 1. [38] In practice there may be problems.

Table 1

Which of the following are the most important factors when selecting a new password? Mark all that apply

Easy to remember	46%
Short and easy to enter	8%
Fun or interesting	9%
Strength (i.e. hard to guess)	71%
Other, please specify	7%

What are the possible ways to guess someone else's password? There are four traditional methods to do that guessing.

The case of default passwords; these may be factory default ones (wifi, switch), or those of a lazy system administrator or user: password, asdfgh, 123456, etc. See, e.g. the case of the French bank, mentioned above. [10]

The case of connection between the login name and password: there is a a formal or a logical or a personal connection between login name and password, For example, login – login19, or romeo – juliet, or Obama – president. See for example the intrusion into Obama's Twitter account above. [7]

The dictionary attack: the attacker collects possible, frequently used passwords into a list and a program tries them one by one at a slower speed (online) or at a higher speed (offline, when shadow passwords are stolen somehow). There are quite a lot of real-life passwords out there that can be obtained easily; see the above mentioned examples and others. This method is useful especially when the attacker wants to crack as many accounts as s/he can – among a lot of people there always will be enough who use simple passwords.

The brute force attack: when the attacker applies a program to try *all* the possible character combinations as passwords.

In short, a "good" password is one when none of the mentioned cracking methods would be successful, or at least not in a reasonable time period.

The ease of remembering our passwords is also not a simple problem. According to the publicly known offline cracking speeds (below) we have to remember quite long passwords. There are techniques which can help you to generate passwords which are easy, or at least easier, to memorize. "The third folk belief is that random passwords are better than those based on mnemonic phrases. However, each appeared to be just as strong as the other. So this belief is debunked. The fourth folk belief is that passwords based on mnemonic phrases are harder to remember than naively selected passwords. However, each appeared to be just as easy to remember as the other. So this belief is debunked.“ [20]

3 Discussion

In this part I am going to point out that some of the widely applied rules are false; first of all, that a strong password ought to contain all kind of characters and must be totally meaningless, perhaps a sequence of random characters. Second, HP published a password generation method that gives a false feeling of security; its weakness is apparent not just today, but it must have been considered as a security risk even when it was published ten years ago.

3.1 The Basic Character Set – Examples

The most important rule nearly everywhere is that passwords must contain all kinds of characters: lower and upper case letters, digits and punctuation marks, or other special characters. This, by itself, is simply not true.

First, let us see two theoretical examples from higher education. "A password based on only small letters, capital letters or numbers has a small key-space. This makes it more easy for brute-force, just because it limits the possibilities." [21]

József Ködmön, associate professor at the University of Debrecen, Hungary, whose research field is cryptography and data security, says that the ideally good password does not contain any meaningful words or expressions and should contain different kind of letters, digits and special characters. [22]

Some examples from the practical life follow. First of all, see Bob Lord's post on Twitter blog, cited above, in the introduction. [1]

The Gmail recommendation: “Use a password with a mix of letters, numbers, and symbols. There are only 26^8 possible permutations for an 8-character password that uses just lowercase letters, while there are 94^8 possible permutations for an 8-character password that uses a combination of mixed-case letters, numbers, and symbols. That's over 6 quadrillion more possible variations for a mixed password, which makes it that much harder for anyone to guess or crack.” [23]

Google Password Help says: “Tips for creating a secure password: Include punctuation marks and/or numbers. Mix capital and lowercase letters. Include similar looking substitutions, such as the number zero for the letter 'O' or '$' for the letter 'S'. Create a unique acronym. Include phonetic replacements, such as 'Luv 2 Laf' for 'Love to Laugh'.” [24]

The Federal Trade Commission (referenced by Bob Lord in Twitter blog) says: “Make your password at least 10 to 12 characters long, and use a mix of letters, numbers, and special characters”. [25]

Twitter: “When you set up your account, be sure to choose a strong password (at least 10 characters that include upper and lower case characters, numbers, and symbols).” [26]

Why is this rule not true? If an attacker could manage the crack with one of the first three guessing methods, the structure of the password would not be interesting, or in other words, the user was careless. So let us suppose the case of a brute force attack. In this case we can be sure that the attacker will find the password – if s/he has enough time. This means that we must take into consideration the number of the different character combinations that the attacker must try in order to find the real password and the speed s/he can provide for the attack.

If one increases the number of the elements in the basic character set, the number of all the possible combinations can be calculated by a power function (x^a, where *x* is the number of the possible characters and *a* is the length of the password). Instead of this we can increase the length of our password, which means that the number of the combinations will be determined by an exponential function (a^x). It is well known that exponential functions shoot up significantly quicker than power functions. So length is more important than the basic character set.

It seems that people usually do not apply the rule of the mixed character set, and it is not as if it was generally known that the length is the more important parameter. See the rockyou.com password structure analyses below.

3.2 Method Recommended by HP

In 2003, HP published and recommended a password generating method to produce different passwords for different sites based on a single and simple password provided by the user. A small program concatenates the single password and the site name, then calculates the MD5 hash value, converts it into ASCII by base64 and truncates that to 12 characters. “1. The algorithm cannot be inverted to discover the user password even if the site name is known. 2. The algorithm is a standard, meaning any implementation must produce the same output for a given input. 3. It is highly unlikely that two different inputs will produce the same output. ... In this example, the unguessable password is qwerty.” [27] The software utility can be downloaded from HP[1] even today.

At first glance it is a big idea; the user can select one easy-to-remember basic word as the password, while the result is different, long and random-like passwords for each account. It *is* marvelous, as long as nobody knows that the user uses this method. But what if the attacker can get some foreknowledge about the user's password generating method, e.g. s/he catches a glimpse of the generator program on the user's screen? In this case, the attacker's situation becomes very comfortable: it is enough to use a short list of basic (pass)words, because the password generating algorithm is known and pretty simple.

3.3 Initials of Poems or Long Sentences

A frequently recommended method is to select a part of one of your favourite poems and use the initials of the words or those of the lines as a password. Ködmön also advises this, with some additions. [22] Computersigh.com also recommends this method, with the extension of adding some complexity by changing some of the letters to upper case ones and inserting some digits [28] or using similar looking substitutes (zero and o, numeric one and letter l etc.). [24] Using this method to generate your own passwords, you should prefer your own sentences to classical poems. Having so many electronic libraries it would not take a huge amount of energy to create a list of possible passwords from initials of the best-known poems. Another method [28] suggests: Start with a sentence or two about ten words total, e.g. “Long and complex passwords are safest. I keep mine secret.” Using the first letter of every word, turn your sentences into an acronym: “lacpasikms” (10 chars). Add complexity, make only the letters in the first half of the alphabet uppercase: “LACpAsIKMs”. Add length with numbers: “LACpAs56IKMs”. Add more length with punctuation and/or symbols: “?LACpAsIKMs)”. This looks like a 14-character-long random password (~10^{27} combinations). It can be done, but I think that most people would not like a password generating method consisting of so many steps.

[1] http://www.hpl.hp.com/personal/Alan_Karp/site_password/index.html

3.4 Random Words with Mnemonic Technique

According to another piece of advice, you ought to select some, e.g. four, words randomly, concatenate them to a single password string and use some mnemonic technique to memorize it. For example, let the four words be 'correct', 'horse', 'battery' and 'staple'. The password would be 'correcthorsebatterystaple'; then try to imagine as if a horse said to you, “that is a battery staple” and then you answered “that's correct”. [29]

Correcthorsebatterystaple.net runs a password generator (built and maintained by Afterlight Web Development) on this basis. The default settings are: four words, a minimum length of 15 characters, the separator is the hyphen, append a random single digit at the end. Considering that the basic dictionary contains 10,000 (English) words, we can calculate the number of possible password combinations: 10^{17}, the same as about the possible combinations of an 8.7-character-long, random-like password. Using some different separators or changing the initials to uppercase would give us a result of about $4*10^{20}$, and this is the same as the number of the possible password combinations of a 10.5-character-long random string. These results do not seem to be very good, seeing the time we would need to crack such passwords.

3.5 Resources Needed to Crack a Password

The difficulty of guessing simply means: how much time would be needed to crack a password? In such a case, “resource” has a complex meaning: the hardware, software and/or any kind of background knowledge, included but not limited to, possessing the shadow passwords. Every piece of background information about the password(s) to be cracked may help the attacker a lot, perhaps too much.

There exist two very different possibilities to crack passwords: we can speak about online and offline password cracking. Online cracking means that the attacker tries to log in to the system using the username s/he wants to crack and tries different passwords. In this case, if the system is run properly, s/he has the possibility of a very limited number of tries and/or time. Any system that has been properly set up will not let you try an unlimited number of passwords, especially not at high speed. This means that an attacker can try a few dozen or a maximum of a few hundred passwords per second.

Offline cracking means that the attacker could somehow manage to get the shadow password(s) and tries to crack them using his/her own resources, of which the computational speed depends on only the hardware-software configuration s/he has.

Let us see the methods an attacker can use against our personal passwords (or those of our company) and their time consumptions.

3.5.1 Default Passwords

Default password mean not only factory default ones (e.g. in some wifi equipment) but those of the lazy system administrators and users. For example: password, password01, asdfgh, 123456, secret, letmein, etc. Use 'top password' in Google for more examples. Every attacker will try the most common passwords as the first step; see e.g. the case of the intrusion into the French bank mentioned above. [10] Using these generic easy-to-guess passwords is an invitation to being hacked, as if you left the starter key in the door of your car. These need only an infinitesimal amount of time.

3.5.2 Logical or Literal Connection

Another common error users may commit is when they select a password that is in connection with the login name or the person. So, the well known rule is, do not use a password such as your birthdate, the names of your children, your phone number, etc. In general, do not use anything that is in connection with your person, especially if this is a public data element. See the case of Obama's Twitter account mentioned above [7], and many other similar ones.

Also, never use passwords that are simple derivations of your login name, e.g. bob – bob12, bob – bob!bob etc. These simple derivations can easily be generated.

These kind of passwords, or at least a carefully selected subset of them, may be used even in an online attack. Using these kind of passwords is also serious carelessness. These also need only an infinitesimal amount of time, too.

3.5.3 Simple Dictionary Attack

The next step of the attacker is the simple dictionary attack. S/he collects the most probable passwords into a list, called a dictionary, and then applies a software tool to check them one by one. This means that it is highly recommended not to use any dictionary words, or more generally, do not use any words or expressions for which Google would return any results (and also do not use a word you entered into Google previously).

A dictionary attack is always applied before the full brute force attack, simply due to the fact that even the largest dictionaries will contain far fewer words than a brute force attack must try. In the case of a system that is run normally, a traditional dictionary attack cannot be used online, because there must be some time delay between the failed login attempts, and too many failed attempts will trigger a security alert for the system administrator.

An offline dictionary attack needs little time, as a dictionary contains only a very limited number of words compared to the brute force attack.

3.5.4 Brute Force Attack

The brute force attack means that a program will try all the mathematically possible character combinations.

This kind of attack may be performed only in offline mode, i.e. when the attacker has succeeded in getting the shadow password(s). In this case we can be sure that the password will be revealed, and the only question is this: How much time would it need? In other words: How many tries can be performed in a second and how many combinations has to be tried?

In 2009 commercial products were available that claimed the ability to test up to 2,800,000,000 NTLM passwords per second on a standard desktop computer using a high-end graphics processor. [30]

At the Passwords^12 Conference in Oslo at the end of 2012, Jeremi Gosney demonstrated extreme cracking speeds with a Virtual OpenCL (VCL) that was running the HashCat password cracking program across five servers equipped with 25 AMD Radeon GPUs and communicating at 10 Gbps over Infiniband switched fabric. He could provide an unbelievable 348 billion tries/sec (NTLM password hashes), which means that a 14 character long WinXP password, for example, could be cracked just in six minutes (see Table 2).

Table 2

Gosney's cracking speed

algorythm	*tries/sec*
NTLM	348 billion
MD5	180 billion
SHA1	63 billion
LM	20 billion
bcrypt (05)	71,000
sha512crypt	364,000

Gosney's team was at a point where their implementation of HashCat on VCL could be scaled up to supporting even 128 AMD GPUs. [18] [19]

So this means that the last tool, brute force cracking, can work at a cracking speed of between some thousands and 350 billion tries per sec, depending on the resources the attacker has and the hash function the system uses to calculate the shadow passwords.

3.5.5 Advanced Dictionary Attack

Because brute force attacks may need huge amounts of resources, attackers may want to reduce their efforts, of course. On the other hand, users may want more secure passwords, or at least passwords they think are more secure, without using random strings. There are some well-known methods to do that, so attackers knowing these methods may be able to take advantage of them to optimize their cracking efforts. This results in a method somewhere between the traditional dictionary and brute force attacks, and this could be called as an advanced dictionary attack.

An old, well-known trick is to use similar looking substitutions in the original plain text of the selected password, e.g. password – p@ssw0rd. This method is so well known that it has its own name ("leet") as an alternative alphabet. Because it is so well known we must suppose an attacker would try this. This method was used in empirical research as well, "...for each word from a dictionary file … make common number substitutions, such a 1 for I, 5 for S etc." [20] So converting a simple password to leet alphabet does not seem to be a good idea.

There may be other password generating habits, and analysing password structures may help attackers a lot. (Un)fortunately, a very large number of passwords has been revealed (of which I referenced some cases above), so would-be attackers have more than enough ammunition to determine the most common password structures. And we, as well, can do some analyses on revealed passwords to see if there are typical password structures or not, and, if yes, what the most common password structures are.

As described above [14], rockyou.com was cracked in 2009 and about 32 million passwords were made public. I downloaded the list of unique passwords [31] to do some structure analyses. The list contained 14,344,391 unique passwords. There is no formal proof, of course, that this list (or any list) is really the original password list. Only the system administrator could have confirmed the originality of the password list, and only if nobody had changed their passwords between the theft and the sysadmin's confirmation. However, it is *said to be* a real password list and it looks something like that. After some cleaning, i.e. removing lines which seemed to be converting errors (too long lines containing html codes), 14,342,415 items remained, of which the length of 1,789 items is longer than 32 characters. It is not impossible to apply such long passwords, especially if one uses a password manager or copy-and-paste.

Some elementary statistical data follows in Table 3, while Figure 1 shows the most common password lengths.

Table 3

rockyou.com password statistics

average password length	8.74
length <8 characters	33.00%
8 <= length <= 12 characters	59.90%
length >12 characters	7.10%
length >=10 characters	31.03%

The minimum password length is 1. The average password length is 8.7, which might have been just enough in 2009 but is surely not enough today. One third of the passwords were not long enough even at that time. On the other hand, another one third of the passwords had quite a good length of at least 10 characters.

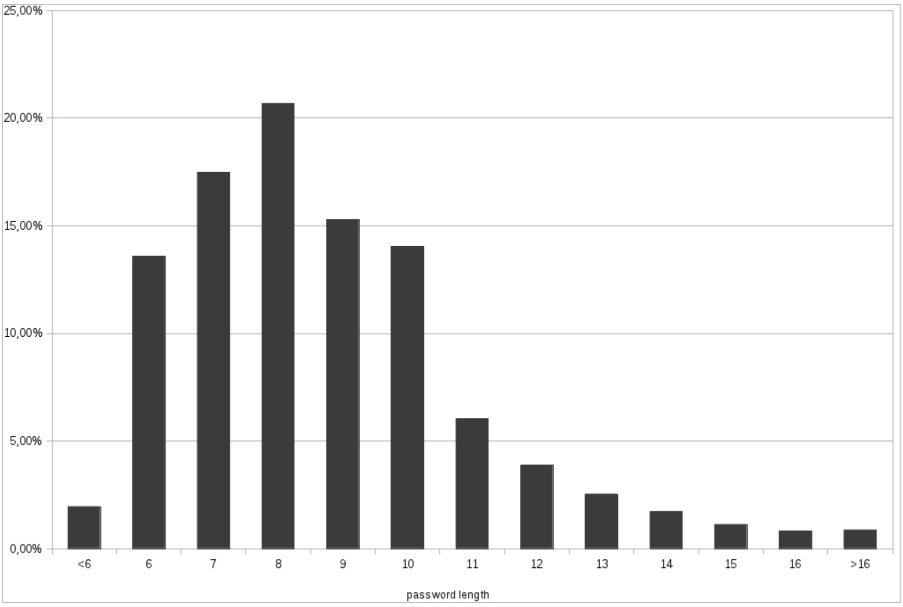


Figure 1
Most common password lengths

I performed some pattern analyses to determine the most common password structures. The character groups I searched for are these: lower case letters, upper case letters, digits, punctuation marks, space, other or special characters.

In the first step, I converted the original passwords, substituting their individual characters according to table 4. So “aaaaa00” means 5 lower case letters and 2 digits at the end, e.g.

Table 4
Character substituting

original	*substituted*
a-z	a
A-Z	A
0-9	0
.,_!?/:;"'-	.
space	_
others	@

Some detailed statistics follow in Table 5. More than one quarter of the passwords consist of lower case letters only. Passwords consisting of only uppercase letters or punctuation marks or others are not preferred; their proportion is less than 2%.

Table 5
rockyou.com password statistics

contains only:	
lowercase letters	26.00%
digits	16.40%
uppercase letters	1.60%
punc. &/or spec.	0.04%
contains space	0.48%
contains at least one:	
uppercase	9.31%
digit	68.08%
punct. or spec.	6.62%
lower+digit	42.36%
lower+upper+digit & none other	2.67%
lower+upper+digit+punc/speci	0.03%

More than one quarter of the passwords consist of only lower case letters. Digits are preferred to uppercase letters or others; two third of the passwords contain digits while only about 16% of them contain uppercase letter(s) or punctuation mark(s) or other special character(s). This means that most people do *not* follow the general – false – rule of passwords, that a password must contain all kind of character types.

Table 6 contains the first 20 of the most common password structure patterns and their lengths. These patterns represent more than the half of the whole set (51.52%). 8 patterns out of the 20 are longer than 8 characters. It seems interesting that these patterns consist of only lower case letters, digits or lower case letters and appended digits. No capital letters, no punctuation marks or special characters.

Table 6
rockyou.com password statistics, most common password structures

structure	%	*length*
aaaaaaaa	4.80	8
aaaaaa	4.19	6
aaaaaaa	4.08	7
aaaaaaaaa	3.60	9
0000000	3.40	7
0000000000	3.33	10
00000000	2.99	8
aaaaaa00	2.93	8
aaaaaaaaaa	2.91	10

000000	2.72	6
000000000	2.14	9
aaaaa00	2.04	7
aaaaaaa00	1.91	9
aaaaaaaaaaa	1.87	11
aaaa0000	1.64	8
aaaa00	1.50	6
aaaaaaaa00	1.49	10
aaaaaa0	1.35	7
aaaaaaa0	1.32	8
aaaaaaaaaaaa	1.32	12
total	*51.53*	

Some common passwords of those that consist of only lower case letters are: password (what a surprise!), iloveyou, princess, sunshine, football, superman, zorro, zzzzzzz. Common given names also appear in the list, e.g. michelle, jennifer, etc.

For an attacker it might be a winning strategy to perform a traditional dictionary attack first, applying a word list of 6-12-character-long common lower case words (e.g. monkey, qwerty, nicole, soccer, peanutbutter, sonyericsson, heartbreaker). This would find a lot of the simple passwords among nearly one third (30.06%) of the whole password set. In the second step, s/he would try a word list of six-character-long words with an appended two-digit number (soccer12, summer07, nicole12, etc.).

A brute force attack against shadow passwords ought to be designed like this: in the first step s/he would attack against the 6-10-digit-long numeric passwords. With a cracking speed of one billion tries per second, one would need about ten seconds to finish them (14.59% of the whole password set, nearly every seventh account). Then pattern 'aaaa00' would fall in 0.05 sec and pattern 'aaaa0000' in 4.57 sec (1.5%, 1.64%, respectively).

Then passwords of 'aaaaa00', 'aaaaa00' patterns would need about 32 more seconds (4.97%). At this point, our would-be cracker would need far less than a minute to crack more than one fifth of the accounts (22.69%).

Then pattern 'aaaaaaa00' would follow (a bit more than 13 minutes, 1.91%), then the lower case passwords of 6-9 characters in length ought to follow; and they would capitulate in about 1.5 hours (an additional 16.67% of the accounts).

This means that far less than 2 hours would be enough to crack nearly four-tenth’s (39.37%) of the whole unique password set (more than five and a half million). This seems to be quite effective; there is a chance that the theft of the shadow passwords has not been discovered yet. Longer and/or complicated passwords would need far more time to crack, e.g. the passwords consisting of ten lower case characters would need 1.6 days.

The efficiency of a brute force attack seems to be characterized by a logarithmic-shaped function curve. It shoots up suddenly (with the weakest passwords) then stabilizes (see the Philips case above). As Ducklin says in [37]: “don't be at the left hand side of the graph.”

Supposing that people's password selecting habits change slowly in general, I would say that an attacker could build up quite an efficient cracking strategy by cracking about 40% of the passwords in less than 2 hours.

You can find a much more detailed structure analysis method in [32]. One may be interested in that, especially if s/he wanted to design a very efficient and scalable password cracking software in general.

As for us, we are interested in generating passwords hard to crack, so we need to know the most common structure patterns included in table 6 to avoid them.

3.6 How to Create a Good Password

As “good” passwords must be both easy-to-remember and hard-to-guess, we can formulate some basic rules and methods to help us create “good” passwords. First of all, as you may know best what kind of things you can memorize, use your own method, but keep in mind the possible approaches an attacker might use against your password. Never think that it cannot happen to you. Below some tips follow based on what we have seen above.

3.6.1 Possible Methods

Use a combination of words and some extensions, like computersigh.com's tip above, but use your own method to create your password, a method that is convenient for you and not known to others. Keep in mind that the length of the password is more important than the basic character set it uses, and the structure itself need not be too complicated. For example, as we know, Shakespeare was born in 1564 in Stratford-upon-Avon, so let us combine these together: Shakespeare1was5born6in4Avon. 29 characters long (30, if the dot is part of the password), not a common structure and easier to remember than computersigh.com's password.

You may also use meaningless words of non-existing languages that sound great at least to you. Drioliano_rodiatenno! – 22 chars.

Prefer foreign, and not English, languages when selecting basic words. Dictionary-based attacks are language dependent. If you have a national keyboard use rare national characters, too, but test it before.

Try some keyboard patterns (but not 'asdfgh' or others like that, of course).

3.6.2 Minimal Password Length

What is a reasonable length for a password in general? You can calculate it easily. Supposing that you have a good enough password which could only be cracked by brute force, calculate the number of possible combinations (the number of characters in the basic character set lifted up onto the power of the length), then divide that number by the supposed cracking speed (measured in tries/sec), then by (60*60*24*365) and you will have the result in years. The bigger the result, the better security you have. Keeping in mind Gosney's results, let the hypothetical cracking speed be 10^{12} tries/sec. Because we can never know... let's multiply it by one thousand and calculate with a value of 10^{15} tries/sec.

For example, see our second example, Drioliano_rodiatenno!. Lower case letters, upper case letters and some others, no digits; the basic character set consists of about 60 characters. As 60^{22}~10^{39}, $10^{39}/10^{15}$/(60*60*24*365)~10^{19}, so there are about 10^{39} possibilities to try, which would need about 10^{19} years. This looks quite good, especially considering the fact that our Universe is supposed to be ~10^{10} years old. And it seems to be not too hard to memorize. A similar password, only 12 characters long, would give us a result of about 70 years. So I suggest that never use shorter passwords and keep in mind: though unbelievable, Gosney's cracking speed is the latest *public* result only.

And do not think it needs too much time to enter such a long password. It needs less time than to unlock your front door, which is a user authentication on the basis the user, i.e. you, has something. You may live in a district where nobody locks their front doors. A computer network is not a world like that.

We suppose at this point that the system for which we use this password is run properly and uses a reasonable hash function to store passwords.

Additional rules of changing your password regularly, etc., must be taken care of, obviously.

3.6.3 Roundabouts, Bypasses

What we discussed above is about how to guess someone else's password or, on the other hand, how to make that too hard. But beware of many kinds of bypasses. You may have the best password in all the world and that will do nothing for you if an attacker, can for example apply a hardware keylogger into the back of your computer (3 secs to install and remove), a small webcam into the ceiling, a trojan horse program into your system, some social hacking, or anything else. See for example the “case study” of the Bastard Operator and the Bank Manager. [32] There is an unimaginably large set of bypasses, so be careful.

3.7 Using and Forcing Good Passwords

In private life it is in our own interest to have strong enough passwords.

In business life it is in our own enterprise's interest. Businesses, first of all, ought to have clear rules, and then make employees keep them. Since any rule is worth only as much as it is kept, businesses should do regular checks. Passwords that the sysadmin can crack with the company’s own resources, which are perhaps very restricted, are really too weak. Online attacks, of course, must be recognized and stopped in time.

We all must keep in mind, especially after Gosney [18] [19] and unlike Denning [4], that if you want security, you have to use physical, administrative and algorithmic protection in parallel. Each chain is only as strong as its weakest link.

3.8 Teaching is Very Important

In our information age it is very important for us to teach not only theoretical material but good practical examples as well. Both the amount of data stored in computers and our dependency on these data increase day by day, so IT security is one of the most important fields in private as well as in business life. As passwords are, and will remain, the first and most important authentication method, we have to concentrate not least on this field.

I, as a teacher, have to draw attention to the fact that the problem shown and discussed above is not simply the problem of only the IT sector, but rather that of education as well. This is not only because passwords are also widely used in education (scholar information systems, e-learning systems, etc.) but because even teenage children may be in danger when they do not know what they are doing when they select their passwords (to log in to gmail, Facebook, etc.). The next generation should also learn the theoretical background. If not, they will not know how passwords work and why it is necessary to select strong passwords or, better to say, what strong passwords are. Students' and children's attention cannot be drawn to such problems too early. If we investigate students' skills and knowledge in the fields of computer sciences here in Hungary or in a wider area, in Central Europe, we find an alarming situation. [34] [35] I, personally, do not think that the situation in the other parts of the world would be significantly better or especially good.

We have to teach our children how to select a reasonably good password and we have at least to try to teach our students that, too. This is not an easy job. A lot of people, unfortunately, prefer laziness to security.

“Results indicate that, in general, users do not vary the complexity of passwords depending on the nature of the site (bank account vs. instant messenger) or change their passwords on any regular basis if it is not required by the site. Users report using lower case letters, numbers or digits, personally meaningful numbers and personally meaningful words when creating passwords, despite the fact that they realize that these methods may not be the most secure.” [33]

The China case cited above and many others show that the human factor is the weakest link in the chain of security. This can be developed only by education, by which we can save not only the enterprise property but the privacy of our young children, too.

Conclusions

The most frequent rule of password selection, i.e. that a password must contain lower case and upper case letters, digits and punctuation marks or other special characters, is not true. The length of passwords is a significantly more important factor than the basic character set. Also, any foreknowledge can highly improve the efficiency of a password cracking attack; so it is recommended not to use typical password patterns and/or generating methods. Really good passwords do exist, and can protect your accounts well. Users are advised to learn what hash function their operating system uses to calculate the shadow passwords. If the hash function is weak, that itself will be a security risk, beyond the users' scope. To improve our security level in general, it is important to teach not only the basic rules but their background as well.

References

[1] Lord, Bob: Keeping our users secure, 1 Feb. 2013 http://blog.twitter.com/2013/02/keeping-our-users-secure.html

[2] Cushman, Reid: Primer Authentication of Identity, Project Health Design ELSI Team, University of Miami, 2007, p. 2

[3] Haley, Kevin: Password Survey Results, Symantec Official Blog, 26 Mar. 2010, http://www.symantec.com/connect/blogs/password-survey-results

[4] Denning, Dorothy: Cryptography and Data Security, Addison-Wesley, 1982, p. 162

[6] Leyden, John: Office workers give away passwords for a cheap pen, The Register, http://www.theregister.co.uk/2003/04/18/office_workers_give_away_passwords/, 18 April 2003

[7] Mesquita, Rafael: Frenchman convicted for hacking Obama http://www.boston. com/business/technology/articles/2010/06/25/frenchman_ convicted_ for_hacking_twitter/, 25 June 2010

[8] Aspan – Baldwin: Sony breach could cost card lenders $300 million, Reuters, http://www.reuters.com/article/2011/04/29/us-sony- creditcards-cost-idUSTRE73S0FL20110429, 28 April 2011

[9] Siegler, Mg.: One Of The 32 Million With A RockYou Account? You May Want To Change All Your Passwords. Like Now., http://techcrunch.com/2009/12/14/rockyou-hacked/, 14 Dec. 2009

[10] Weitzenkorn, Ben: Bank of France's Accidental Hacker Acquitted, http://www.technewsdaily.com/8140-accidental-hacker-bank-france.html, 21 Sep. 2012

[11] Gee – Kim: http://godaigroup.net/publications/doppelganger-domains/, 6 Sep. 2011

[13] Calin, Bogdan: Statistics from 10,000 leaked Hotmail passwords, Acunetix Web Application Security, http://www.acunetix.com/blog/news/statistics-from-10000-leaked-hotmail-passwords/, 6 Oct. 2009

[14] Consumer Password Worst Practices by The Imperva Application Defense Center (ADC), 2010, http://www.imperva.com/docs/wp_consumer_ pass word_worst_practices.pdf

[15] Yang - Hung – Lin: Loose Password Security in Chinese Cyber World Left the Front Door Wide Open to Hackers – An Analytic View, ICEC 12 Proceedings of the 14th Annual International Conference on Electronic Commerce, ACM, 2012. pp. 121-126

[16] Nilsson, Anders: Statistics of "450.000 leaked Yahoo accounts", http://pastebin.com/2D6bHGTa, 13 July 2012

[18] Update: New 25 GPU Monster Devours Passwords In Seconds, http://securityledger.com/new-25-gpu-monster-devours-passwords-in-seconds/, 4 Dec. 2012

[19] New 25-GPU Monster Devours Strong Passwords In Minutes, http://it.slashdot.org/story/12/12/05/0623215/new-25-gpu-monster-devours-strong-passwords-in-minutes, 5 Dec. 2012

[20] Yan - Blackwell - Anderson – Grant: The memorability and security of passwords: empirical results, Security & Privacy, IEEE, Vol. 2, Issue 5, Sept.-Oct. 2004, pp. 25-31

[21] Bakker – Jagt: GPU-based password cracking, University of Amsterdam, System and Network Engineering, 2010, p. 7)

[22] Ködmön, József: Biztonságosabb felhasználóazonosítás az egészségügyben [More secure user authorization in health care], IME - Az egészségügyi vezetők szaklapja [Journal for Managers in Health Care], Vol. 6, Issue 9, Nov. 2007, pp. 46-51

[23] The Gmail Team: Choosing a smart password, http://gmailblog.blogspot.hu/2009/10/choosing-smart-password.html, 7 Oct. 2009, downloaded 5 Feb. 2013

[24] https://accounts.google.com/PasswordHelp, no date of uploading, downloaded 9 Feb 2013

[25] Make Computer Security One of Your New Year's Resolutions, http://www.consumer.ftc.gov/blog/make-computer-security-one-your-new-years-resolutions, 3 Jan. 2013

[26] Keeping Your Account Secure, https://support.twitter.com/articles/76036-keeping-your-account-secure#, 5 Feb. 2013

[27] Karp, Alan H.: Site-Specific Passwords, Hewlett-Packard Company, 2003

[28] Password statistics, http://computersight.com/communicationnetworks/security/password-statistics/, 16 Feb. 2011

[29] http://correcthorsebatterystaple.net/, 12 Feb. 2013

[30] Belenko, Andrei: Password Recovery SolutionsForensics & Investigation IT- Security Audit, Elcomsoft Proactive Software, 16 March 2009

[31] http://www.skullsecurity.org/wiki/index.php/Passwords
downloaded 17 Oct 2012

[32] The Bastard Operator From Hell #7 http://bofh.ntk.net/BOFH/0000/bastard07.php

[33] Shannon, Riley: Password Security: What Users Know and What They Actually Do, Wichita State University - Software Usability Research Lab, February 2006, Vol. 8, Issue 1)

[34] Kiss, G.: Measuring Computer Science Knowledge Level of Hungarian Students specialized in Informatics with Romanian Students attending a Science Course or a Mathematics-Informatics Course / TOJET: The Turkish Online Journal of Education Technology, Volume 11, Issue 4, ISSN: 2146 – 7242, pp. 222-235

[35] Kiss, G.: Comparison of the Programming Knowledge of Slovakian and Hungarian Students / Procedia of Social and Behavioral Science Journal különszám, ISSN: 1877-0428, p. 10

[36] Florencio – Herley: A Large-Scale Study of Web Password Habits, WWW 2007 / Track: Security, Privacy, Reliability, and Ethics, ACM, pp. 657-665

[37] Ducklin, Paul: Cracking passwords from the Philips hack - an important lesson, http://nakedsecurity.sophos.com/2012/08/22/cracking-passwords-from-the-philips-hack/, 22 Aug. 2012

[38] khaley, symantec employee: Living with Passwords, http://www.symantec.com/connect/blogs/living-passwords, 25 March 2010

Multicriteria Cruise Control Design Considering Geographic and Traffic Conditions

Balázs Németh, Alfréd Csikós, Péter Gáspár, István Varga

Systems and Control Laboratory, Computer and Automation Research Institute,
Hungarian Academy of Sciences, Kende u. 13-17, H-1111 Budapest, Hungary
E-mail: bnemeth@sztaki.mta.hu, csikos@sztaki.mta.hu, ivarga@sztaki.hu,
gaspar@sztaki.hu

Abstract: The paper presents the design of cruise control systems considering road and traffic information during the design of speed trajectories. Several factors are considered such as road inclinations, traffic lights, preceding vehicles, speed limits, engine emissions and travel times. The purpose of speed design is to reduce longitudinal energy, fuel consumption and engine emissions without a significant increase in travel time. The signals obtained from the road and traffic are handled jointly with the dynamic equations of the vehicle and built into the control design of reference speed. A robust H_∞ control is designed to achieve the speed of the cruise control, guaranteeing the robustness of the system against disturbances and uncertainties.

Keywords: look-ahead control; multicriteria optimization; robust H_∞ control

1 Introduction and Motivation

The driveline system plays an important role in energy consumption and the emission of vehicles, therefore the development of longitudinal control systems is in the focus of research and the industry. Adaptive Cruise Control (ACC) systems guarantee the adaptation of the vehicle to the environment, such as instantaneous road disturbances, road slopes, rolling resistances, and the speed of preceding vehicles. However, these systems are not able to take into consideration the road and traffic information expected from the subsequent road sections, such as speed limits and road inclinations.

In this paper a longitudinal control system is proposed which is able to consider predicted road and traffic information in the design of the longitudinal control force. Using the signals of road inclinations and speed limits, fuel consumption, the energy required by the actuators, and engine emissions can be reduced. Moreover, the unnecessary activation of the brakes is also avoided, which is desirable for reducing the wear of the brake pads/discs and the loss of kinetic energy. The control of longitudinal dynamics requires the integration of these control components, see [6].

Several methods have already been published on the topic of look-ahead control; see [5, 10]. The robust H_∞ control design method was proposed by [9] for the design of vehicle speed based on road inclinations. In [8] the emission of the vehicle was also taken into consideration. The optimization problem was handled in the same manner, using the receding horizon concept on spatial increments in [3, 11]. The terrain and traffic flow were modeled stochastically using a Markov chain model in [7]. [4] evaluated the approach in real experiments.

This paper focuses on the design of vehicle speed based on signals obtained from the road and traffic. Several aspects are considered, such as road inclinations, traffic lights, preceding vehicles, speed limits, travel times and the effect of engine emissions. Since the designs for different aspects result in different solutions, a balance needs to be achieved between them by using multi-objective optimization. The novelty of the paper is that it considers the signals of traffic lights during the design of speed trajectories. In this way it is possible to reduce the number of unnecessary brakings, accelerations and stop-and-starts, which may considerably increase the required energy, fuel consumption and engine emissions. Since the proposed method also handles speed limits and preceding vehicles, it can be applied on motorways and in urban traffic networks as well.

The paper is organized as follows: Section 2 presents the aspects of the speed design, such as road inclinations, emissions and oncoming road intersections. The design of the control strategy for oncoming traffic lights is detailed in Section 3. Section 4 presents the multi-criteria optimization of vehicle cruise control by the appropriate choice of prediction weights. Section 5 shows the operation of the control system on a transportational route. Finally, Section 6 summarizes the conclusion remarks.

2 Geographic and Traffic Criteria of Speed Design

2.1 Speed Design for Road Inclinations

The design of the speed trajectory is based on road inclinations and speed limits. Since the design of optimal speed has already been proposed in an earlier paper, only a brief summary is given, for details see [9].

Road inclinations are considered in the design of the longitudinal control force. On a downhill slope the speed of an undriven vehicle increases by itself, thus the control force of the vehicle before the slope may decrease. Consequently, the brake system can be activated later, or it is not necessary to be activated at all.

Before the section where a speed limit is imposed, the speed can be reduced, therefore less braking energy is necessary for the vehicle. By choosing the appropriate speed profile according to the road and traffic information, the number of unnecessary longitudinal interventions and their durations can be significantly reduced.

For the consideration of predicted information, the route of the vehicle is divided into *n* sections using *n+1* number of points. The division of the route is of arbitrary lengths. The rates of the inclinations of the road and the locations of the speed limits are assumed to be known at the endpoints of each section. In each section point of the road, reference speeds are defined, which depend on the speed limits. The speed at section point *j* should reach the predefined reference speed $v_{ref,j}^2 \quad j \in [1,n]$. The control task is then to track the momentary value of the speed, which is formulated in the following form: $\dot{\xi}_0^2 \to v_{ref,0}^2$.

The road sections to be taken into consideration are qualified by different weights. A weight *Q* is applied to the initial speed and weights $\gamma_i, i \in [1,n]$ are applied to the further reference speeds. A weight *W* represents the tracking of the speed of the preceding vehicle v_{lead} in order to avoid a collision, see [9]. The safety distance between the vehicles is determined according to directives: $d_{st} = 0.1\dot{\xi}_0 + \dot{\xi}_0^2/150$. The prediction weights should sum up to one, i.e., $W + Q + \sum_{i=1}^{n} \gamma_i = 1$. The interpretation of the importance of W, Q, γ_i prediction weights is the following. If *Q* weight is set to 1 and the other weights are 0, the simple cruise control is achieved. If equal weights for Q and γ are set and *W* is 0 in the cruise control, the predicted road sections have the same importance. When *W=1* and $Q = \gamma_i = 0, i \in [1,n]$, only the tracking of the leader vehicle is realized. The optimal determination of weights is a key issue, i.e., to achieve a balance among speed limits, effect of road slopes and traffic situations.

During the design of the vehicle speed the prediction weights are taken into consideration. The momentary vehicle speed $\dot{\xi}_0$ must be modified in the following way:

$$\lambda = \sqrt{\vartheta - 2s_1(1 - Q - W)(\ddot{\xi}_0 + g sin\alpha)} \tag{1}$$

where the value ϑ depends on the predicted road slopes, the reference speeds and the prediction weights:

$$\vartheta = Wv_{lead}^2 + Qv_{ref,0}^2 + \sum_{i=1}^{n} \gamma_i v_{ref,i}^2 + \frac{2}{m}(1 - Q - W)\sum_{i=1}^{n}(s_i F_{di,r} \sum_{j=i}^{n} \gamma_j) \tag{2}$$

where v_{ref} and v_{lead} denote the reference speed and the velocity of the leading vehicle respectively, s_i denotes the length of segment *i* and $F_{di,r}$ denotes the unmeasured longitudinal disturbances.

The calculation of λ requires the measurement of the longitudinal acceleration $\ddot{\xi}_0$. Consequently, the aim of the control design is to track the calculated speed trajectory: $\dot{\xi}_0 \rightarrow \lambda$.

2.2 Speed Design for Emissions

The pollution emerging from road traffic has become a serious environmental issue over the past decades. Modeling the amount and composition of exhaust gases is essential for an effective control aimed at minimizing emissions and fuel consumption. When individual vehicles are analyzed, emission models can be classified into two categories based on the number of input variables: traffic situation models and average speed models. Input variables of the former models include information of the current traffic situation or more specifically, instantaneous acceleration in addition to the speed variable. Average speed models are used if no information is available on the current driving pattern apart from average speed, and thus the output of the model is the emission assigned to validated measurement cycles of the average speed value.

Emission can be described by its temporal rate (emission rate function) or – throughout a journey – by its spatial rate (emission factor function). Emission factors from road vehicles can be derived via different approaches. The functions of single vehicles can be measured by dynamometer tests, by transient chassis dynamometers or by engine emission measurements [14]. The most common approach to obtain emission factors of great resolution with regard to vehicle technology is to sample a range of cars of a given category and emission control technology (e.g. gasoline passenger car 1.4-2.0l, Euro 4), drive them according to pre-defined driving patterns (driving cycles) on a chassis dynamometer, and then record their emissions over such conditions. The total produced emissions divided by the total distance driven results in a mean emission factor which is considered representative of the particular vehicle technology (provided that the vehicle sample is sufficient large) when driven under similar driving conditions as those covered by the driving cycle. Model functions are then fitted to these data sets. A standard method is that emission factors of the pollutants are modeled by convex rational functions of average vehicle speed, see e.g. the model COPERT [13]. For the use of model COPERT in a macroscopic traffic emission framework, see [15]. The emission factor functions are specific for different vehicle classes, fuel types, Euro norms and engine capacities. For vehicle type *c* and pollutant *p*:

$$ef^{\,p,c} = (\sum_{i=0}^{m} \beta_i^{p,c} \dot{\xi}(t)) / (\sum_{i=0}^{n} \delta_i^{p,c} \dot{\xi}(t)) \tag{3}$$

where $\dot{\xi}$ (t) denotes the instantaneous longitudinal vehicle speed, and $\beta_i^{p,c}$, $\delta_i^{p,c}$ are constant model parameters, depending on pollutant *p* and vehicle class *c*.

The following pollutants were modeled in the control design: CO, CO_2, NO_x and hydrocarbons (HC). These are considered the most significant exhaust gases that cause both global (greenhouse effect) and local harms (health problems, acid rain). Elaborating the reaction stochiometrics of internal combustion engines, a linear connection between the fuel consumption and the CO_2 emission of a vehicle can be stated [12]: $ef^{CO_2,c} = K \cdot f^c$, where f^c is the fuel consumption of a type *c* vehicle and *K* depends on the fuel type, e.g. in Diesel fuel *K=26.29*. Unfortunately, further analytic relationships cannot be drawn among emission functions of the pollutants as secondary reactions of internal combustion engines depend on several factors (i.e. engine and fuel type, engine load, technology of engine etc).

During the performance analysis, the normed sum of emissions is examined:

$$ef_{total}(t) = \sum_{c=1}^{N_c} \frac{ef^{CO_2}(t)}{ef_{nom}^{CO_2}} + \sum_{c=1}^{N_c} \frac{ef^{CO}(t)}{ef_{nom}^{CO}} + \sum_{c=1}^{N_c} \frac{ef^{HC}(t)}{ef_{nom}^{HC}} + \sum_{c=1}^{N_c} \frac{ef^{NO_X}(t)}{ef_{nom}^{NO_X}} \qquad (4)$$

where $ef_{nom}^{p} = max_{v\in[60,90]}ef^{p}(v)$ denotes the nominal emission for pollutant *p*.

2.3 Speed Design for Oncoming Intersections

In the proposed control, the third criterion is the consideration of oncoming intersections. The proposed system uses traffic signal scheduling information for the design of the reference speed of the controlled vehicles. This proposition involves the matching of traffic and vehicle control for which certain hardware equipment is required. The technology is similar to that used in transit priority detection with the difference that the control intervention is executed in the vehicle only. In order to establish an effective cooperation, both the infrastructure and the vehicle are equipped with communication devices.

Different communication systems using this layout are known: e.g. the detection method using roadside beacons, the GPS-based detection method and the infrared detection method, see [1]. The former two methods rely more on the devices applied to the infrastructure (e.g. roadside beacons). In the infrared detection method, which is widespread in the United States, the vehicle is detected by a standard traffic inductive loop detector (ILD) embedded in the pavement which alerts the signal controller of the approaching HDV. After receiving the entrance request, the traffic signal controller (TSC) sends the timing data of the traffic signal via an infrared emitter (most commonly located on the signal mast arm or span wire). The scheduling information is then received by the IR (Infrared) detector of the vehicle and further used for the control design. The system layout is illustrated in Figure 1.

The main advantage of loop detection systems is that the system is compatible with commonly used loop detectors. It also does not require line-of-sight or visibility, and IR transponders may be set on already installed traffic controllers. The effective range of the system highly depends on the geographical layout of the intersection and may range up to 500m.

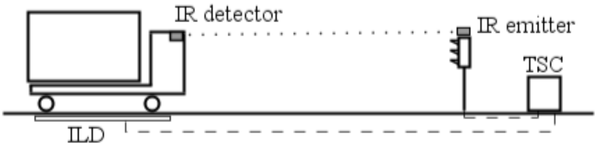


Figure 1
Communication architecture of the transit vehicle system

3 Control Strategy at Traffic Lights

Reference speed (1) is designed by taking road inclinations, preceding vehicles and speed limits into consideration. In the following a traffic signal of the intersection as a further criterion is used in the speed design. It is assumed that the traffic signaling data are available during the design process. For the strategy, necessary traffic information involves the distance between the vehicle and the traffic light s_{int}, the signal of the traffic light and the expiry time of the current signal. According to this information, the control strategy of speed calculation is chosen by making a decision logic, which is illustrated in Figure 2.

When the vehicle receives an information package from a traffic light, it makes a decision based on its current speed, etc. The following scenarios are analyzed in three cases: there is a green signal along the route, and in two additional cases there are red (or amber) signals.

Case 1: In the first case, the vehicle reaches the intersection during the green signal without increasing its speed, i.e., $s_{int}/\dot{\xi}_0 \leq T_{gr}$. However, the speed of the vehicle may be reduced if it turns at the intersection. In this case the speed at the intersection must be modified to a safe cornering speed, thus $\dot{\xi}_0 = v_{int}$. It is achieved by setting the weight $Q = 1$ and $v_{ref,0} = v_{int}$. The condition for this case is:

$$\frac{2s_{int}}{\dot{\xi}_0 + v_{int}} \leq T_{gr} \tag{5}$$

where T_{gr} is the unexpired green time. Here the linear relationship between the initial speed $\dot{\xi}_0$ and the final speed v_{int} is exploited. Note that in straight motion, the speed and the weights are not modified, thus $v_{int} = \dot{\xi}_0$.

Case 2: In the second case, the vehicle reaches the intersection during the green signal if the speed is increased to the maximum allowed speed. In this case, the speed at the intersection must be modified to the original reference speed, thus $\dot{\xi}_0 = v_{ref,0}$. It is achieved by setting the weight *Q*=1. The condition for this case is:

$$\frac{s_{int}}{v_{ref,0}} \le T_{gr} \tag{6}$$

In this scenario, the intersection overwrites the modified reference speed and high acceleration and deceleration are applied.

Case 3: If the vehicle does not reach the intersection during the green signal, the deceleration of the vehicle and a safe stop condition are required. They are achieved by setting the speed $v_{lead} = 0$ and modifying the weight *W* in the following way:

$$W = 1 - \frac{s_{int}^2}{s_{int,max}^2}, \tag{7}$$

where $s_{int,max}$ is the distance between the vehicle and the traffic light when the signal arrives. Thus, in the calculation of speed, the predicted road information becomes less important when the vehicle is approaching the traffic light, and the stopping manoeuver has priority, i.e., $\dot{\xi}_0 \to 0$.

The further scenarios involve situations when the signal is red (or amber).

Case 4: If the signal is red, the time requirement for reaching the intersection at the vehicle's current speed is calculated. If during this time the red signal turns to green in straight motion, the speed and the weights are not modified, thus $v_{int} = \dot{\xi}_0$. However, the speed of vehicle may be reduced if the vehicle turns at the intersection. In this case, the speed at the intersection must be modified to a safe cornering speed, thus $\dot{\xi}_0 = v_{int}$. It is achieved by setting the weight *Q=1* and $v_{ref,0} = v_{int}$. The condition for this scenario is:

$$\frac{2s_{int}}{\dot{\xi}_0 + v_{int}} \ge T_{red} \tag{8}$$

where T_{red} is unexpired red time. Note that at straight motion the speed and the weights are not modified, thus $v_{int} = \dot{\xi}_0$.

Case 5: If the signal is red and the unexpired red time is too long, it is necessary to stop the vehicle. In this scenario *W* is influenced according to (7).

Also note that in the previously formulated decision logic, the other preceding vehicles are ignored. In the case of a preceding vehicle *W* is modified according to [9].

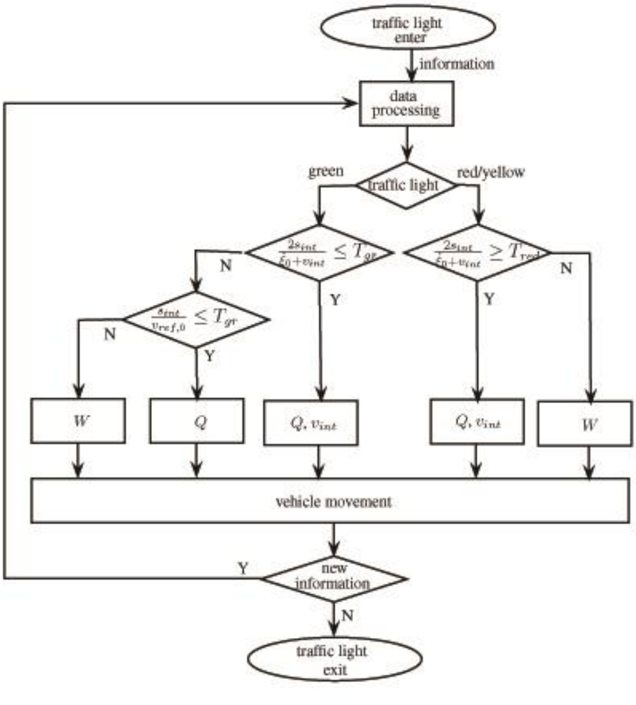


Figure 2

Flowchart of the traffic signal strategy

4 Formulation of Performance Criteria

4.1 Multicriteria Optimization

The aim of this section is to find an optimal speed $\dot{\xi}_0$, which guarantees the joint minimization of the control force, travel times and emission. The fulfilment of these performances individually results in different W, Q, γ_i weights according to equation (1).

In the minimization tasks, W is handled as an exogenous signal, which is set according to different special scenarios, i.e., preceding vehicles or traffic lights, see Section 3. Thus, in the optimization task, the weight W=0 is set in the standard case. During travelling Q and γ_i are calculated and applied. However, in case of preceding vehicles or oncoming intersections they are completed by weight W in such a way that $W+Q+\sum_{i=1}^{n}\gamma_i=1$ is guaranteed.

In the first optimization criterion, the longitudinal force F_{l1} must be minimized, i.e., $|F_{l1}|\to Min$. The force can be expressed as the linear function of prediction weights by using equation (1):

$$F_{l1}(Q,\gamma)=\beta_0(Q)+\beta_1(Q)\gamma_1+\beta_2(Q)\gamma_2+\ldots+\beta_n(Q)\gamma_n \tag{9}$$

where β_i are the coefficients of Q and γ_i. In practice, however, the following optimization form is used because of the simpler numerical computation:

$$F_{l1}^2\to Min \tag{10}$$

The optimal solution leads to $\overline{Q}$ and $\overline{\gamma}_i$, satisfying the constraints $0\le\overline{Q},\overline{\gamma}_i\le1$ and $\overline{Q}+\sum\overline{\gamma}_i=1$. The solution of the optimization problem is found in [2].

The second optimization criterion is the minimization of traveling time. In this case the vehicle has to travel at the predefined reference speed. Therefore, the difference between momentary speed and reference speed needs to be minimized, i.e.,

$$|v_{ref,0}-\dot{\xi}_0|\to Min \tag{11}$$

This means that this optimization criterion is fulfilled if the road inclinations are ignored. The optimal solution of the performance (1) is: $\breve{Q}=1$ and $\breve{\gamma}_i=0, i\in[1,n]$.

The emission model of the vehicle is approximated by using a second order polynomial function according to equation (3): $ef_{total}(t)=\alpha_0+\alpha_1\dot{\xi}_0+\alpha_2\dot{\xi}_0^2$, where $\alpha_0,\alpha_1,\alpha_2$ are constant parameters. There is a formal analogy between $ef_{total}(t)$ and the unmeasured longitudinal disturbances $F_{d1,o}$, see [8]. In the third optimization criterion, the total emission $ef_{total}(t)$ is minimized: $|ef_{total}(t)|\to Min$. In practice, the following optimization form is used:

$$\left(ef_{total}(t)\right)^2\to Min \tag{12}$$

This minimization leads to a quadratic optimization problem, similarly to the first performance. The solutions of the optimization are denoted by $\hat{Q},\hat{\gamma}_i$ weights.

It is important to emphasize that the three performances (minimization of longitudinal force, traveling time or emission) result in different prediction weights. Thus, it is necessary to guarantee a tradeoff between them. In the multi-criteria optimization, three further performance weights are introduced. The roles of these factors are different. Performance weight R_1 is related to the importance of the minimization of the longitudinal control force, performance weight R_2 is related to the minimization of traveling time, while performance weight R_3 is related to the importance of emission. Since there is a constraint on the performance weights, $R_1 + R_2 + R_3 = 1$, a balance between the optimizations tasks can be achieved. The form of the final weights are the following:

$$Q = R_1\overline{Q} + R_2\breve{Q} + R_3\hat{Q} = R_1\overline{Q} + R_2 + \hat{Q}R_3 \tag{13}$$

$$\gamma_i = R_1\bar{\gamma}_i + R_2\breve{\gamma}_i + R_3\hat{\gamma}_i = R_1\bar{\gamma}_i + R_3\hat{\gamma}_i \tag{14}$$

with $i \in [1,n]$. The calculated multi-criteria optimal Q, γ_i prediction weights with the exogenous *W* are used for the calculation of the modified reference speed *λ*, see (1), and they are used during the travelling.

4.2 H∞-based Robust Control Design

During traveling, different disturbances which are not considered in the speed design may influence the vehicle dynamics. Thus it is necessary to find a robust speed controller *K*, which is able to track the calculated speed value. The controller can guarantee robustness against external disturbances, such as sensor noises and road disturbances, and also handle unmodeled disturbances.

The purpose of tracking is to ensure that the system output follows a reference value with an acceptably small error, which is the performance of the system. The control problem is as follows:

$$|\lambda - \dot{\xi}_0| \to Min! \tag{15}$$

where parameter *λ* is the reference value according to (1). Thus, the performance signal is $z = \dot{\xi}_0 - \lambda$.

The standard form of the closed-loop interconnection structure, which includes the feedback structure of the model *P* and controller *K*, is shown in Figure 3.

The control design is based on a weighting strategy. The purpose of weighting function W_p is to define the performance specifications of the control system. In the selection of W_p an accurate matching is required at low frequencies and a less accurate matching is acceptable at higher frequencies. The function W_p is selected as $W_p = \alpha/(Ts+1)$, where α and T are constants. Here, it is required that the steady state value of the tracking error is below *1/α* in steady-state.

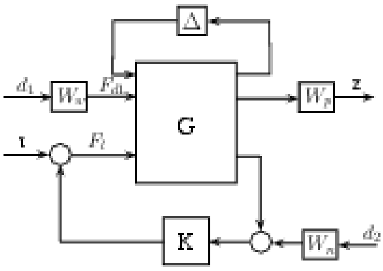


Figure 3
Closed-loop interconnection structure

Three additional weights are also applied. W_n reflects to the speed sensor noise, while W_w represents the effect of longitudinal disturbances. In the modeling, an unstructured uncertainty is modeled by connecting an unknown but bounded perturbation block (Δ) to the plant. The magnitude of multiplicative uncertainty is handled by a weighting function W_u. The weighting functions W_u, W_w and W_n are selected in linear and proportional forms.

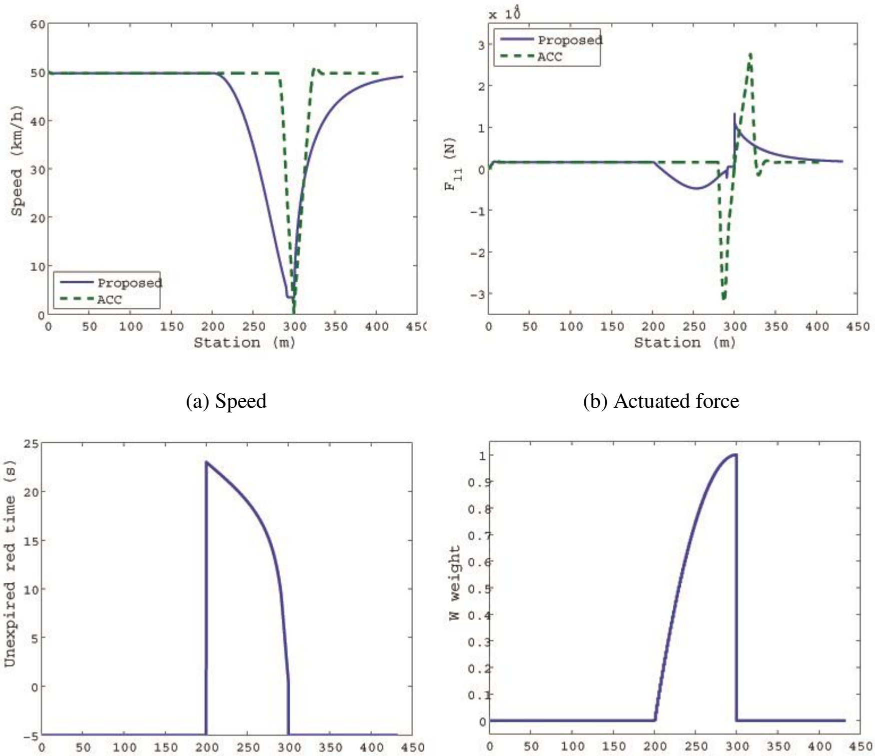
5 Simulation Results

In this section the operation of the vehicle system is analyzed in case studies. Both road and traffic information are taken into consideration. Note that in the simulation example only the longitudinal force will be in the focus. The balance between the three performances are analyzed in another paper, see [8].

In the simulation examples two cruise control systems are compared. The first system uses a conventional adaptive cruise control (ACC) ignoring the predicted weights. This system always tracks predefined $v_{ref,i}$ reference speeds. The system using a cruise control (Proposed) considers the road and traffic conditions through predicted weights. Consequently, this system is able to modify the reference speed during traveling. In the figures the proposed control is denoted by solid line, while the conventional control is denoted by dashed line. Figures show the time responses of the simulation, i.e., the speed, the longitudinal force, the unexpired time and the weight W.

In the first simulation example the vehicle arrives within the range of the traffic light, which is red. Moreover, the expiry time of the red signal is long, thus it is necessary to stop the vehicle (see Case 5 in Section 3). The simulation starts when the distance between the vehicle and the traffic light is 300 m, but the range of the traffic light is 100 m. The proposed control receives the information package of the traffic light at 200 m, therefore vehicle speed is reduced up to this point. The unexpired red time decreases as Figure 4(c) shows. To guarantee the stopping of the vehicle, it is necessary to increase *W* weight, which is shown in Figure 4(d).

The conventional control reduces the speed abruptly, when the vehicle is close enough to the traffic light, see Figure 4(a). Thus, both the duration and the magnitude of the longitudinal force are higher in the ACC case, see Figure 4(b). Less longitudinal force and energy are required during the journey in the proposed control method. The saved longitudinal force is about 16% compared to the conventional cruise control system. Consequently, using the proposed control strategy smaller energy consumption can be achieved.



(a) Speed

(b) Actuated force

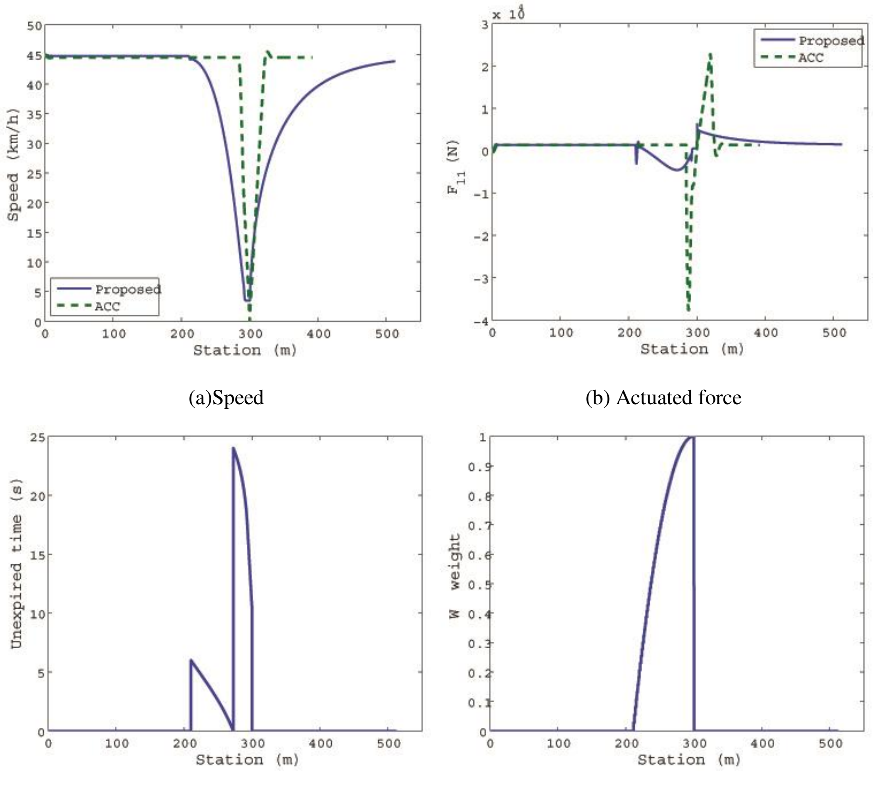
(c) Unexpired red time

(d) *W* weight

Figure 4

Traffic light with long unexpired redtime

In the second simulation example, the vehicle receives the green signal in the range of the traffic light. It shows that during the green signal the vehicle does not reach the intersection; see Case 3 in Section 3. Figure 5(c) shows the unexpired green time and then the red time. Thus, the speed must be reduced. The requirement of the deceleration and the safe stopping at the traffic light is defined by the modification of weight *W*, which is illustrated in Figure 5(d). The speed and the necessary longitudinal force are shown in Figures 5(a) and 5(b), respectively. Less longitudinal force and energy are required during the journey in the proposed control method. The saved longitudinal energy is about 11% compared to the conventional cruise control system.



(a)Speed

(b) Actuated force

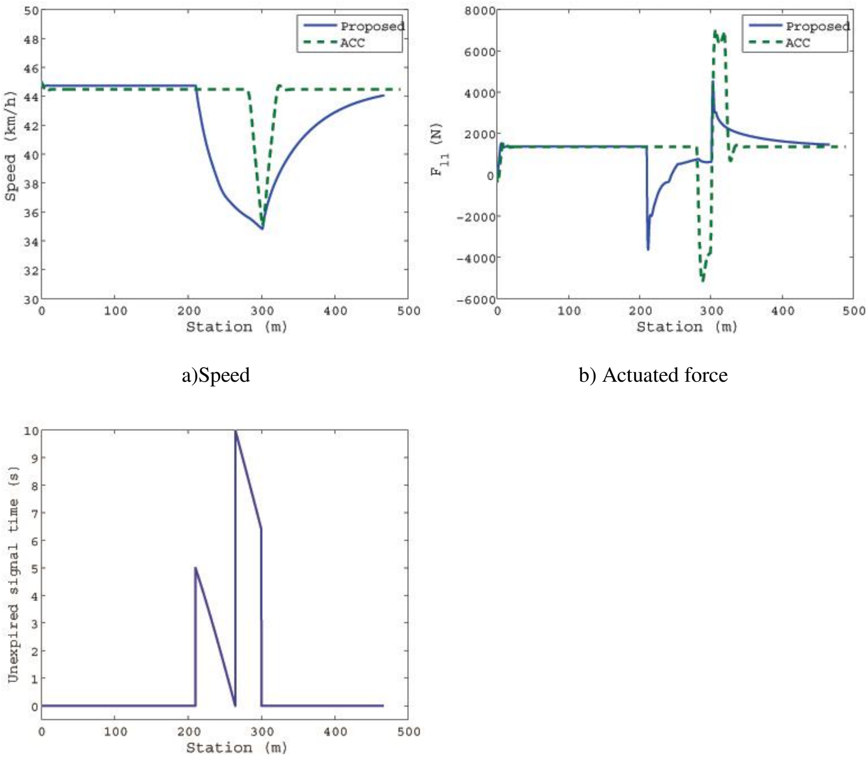
(c)Unexpired red time

(d) *W* weight

Figure 5

Traffic light with green and red signals

In the third example the vehicle receives traffic information about the red signal, which turns to green. Figure 6(c) shows the short unexpired red time, which is followed by the green time. Thus, the speed must be reduced to the safe cornering speed v_{int}, see Figure 6(a). In this scenario v_{int} can be achieved within relatively long time with reduced longitudinal force by exploiting the adhesion coefficient of the road, see Figure 6(b). Less longitudinal force and energy are required during the journey in the proposed control method. The saved longitudinal force is 19% compared to the conventional cruise control system.



a)Speed

b) Actuated force

c) Unexpired red time

Figure 6

ACC systems with a compulsory speed limit

Conclusions

The paper has proposed the design of a cruise control system which is able to exploit information received from both geographic features and traffic. The main result of the research is that an intersection with a traffic light is included in the speed design. An optimal speed trajectory is computed according to the balance between the three factors, i.e., the longitudinal force, traveling time and emission. The control design is based on the robust H_∞ method, in which performance specifications, disturbances and uncertainties are considered. The simulation results show that the designed control reduces the energy required by the actuators.

Acknowledgements

The paper was supported by the János Bólyai Research Scholarship of the Hungarian Academy of Sciences and by the project EITKIC 12-1-2012-0001, supported by the Hungarian Government, managed by the National Development Agency, and financed by the Research and Technology Innovation Fund.

References

[1] K. Gardner, C.D. Souza, Nick Hounsell, Birendra Shrestha, and David Bretherton. Review of Bus Priority at Traffic Signals around the World. *UITP Working Group Technical Report*, 2009

[2] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, London UK, 1981

[3] Erik Hellström, Jan Åslund, and Lars Nielsen. Horizon Length and Fuel Equivalents for Fuel-Optimal Look-Ahead Control. Munich, 2010

[4] E. Hellström, M. Ivarsson, J. Åslund, and L. Nielsen. Look-Ahead Control for Heavy Trucks to Minimize Trip Time and Fuel Consumption. *Control Engineering Practice*, 17(2):245-254, 2009

[5] M. Ivarsson, J. Åslund, and L. Nielsen. Look Ahead Control - Consequences of a Non-Linear Fuel Map on Truck Fuel Consumption. *Proc. Institution of Mechanical Engineers, Journal of Automobile Engineering*, 223:1223-1238, 2009

[6] U. Kiencke and L. Nielsen. *Automotive Control Systems for Engine, Driveline and Vehicle*. Springer, 2000

[7] I. V. Kolmanovsky and D. P. Filev. Stochastic Optimal Control of Systems with Soft Constraints and Opportunities for Automotive Applications. *IEEE Conference on Control Applications, St. Petersburg*, 2009

[8] B. Németh, A. Csikós, I. Varga, and P. Gáspár. Design of Platoon Velocity-based on Multi-Criteria Optimization. *7th IFAC Symposium on Robust Control Design (ROCOND)*, 2012

[9] B. Németh and P. Gáspár. Road Inclinations in the Design of LPV-based Adaptive Cruise Control. *18th IFAC World Congress*, 2011

[10] L. Nouveliere, M. Braci, L. Menhour, and H. T. Luu. Fuel Consumption Optimization for a City Bus. *UKACC Control Conference*, 2008

[11] B. Passenberg, P. Kock, and O. Stursberg. Combined Time and Fuel Optimal Driving of Trucks Based on a Hybrid Model. *European Control Conference, Budapest*, 2009

[12] Abhishek Tiwary and Jeremy Colls. *Air Pollution. Measurement, Modelling and Mitigation. Third edition*. Taylor and Francis Group, Routledge, 2010

[13] M. Ekström, Å. Sjödin, K. Andreasson: Evaluation of the Copert III Emission Model with On-Road Optical Remote Sensing Measurements. Atmospheric Environment, Vol. 38, Issue 38, December 2004, pp. 6631-6641

[14] Thomas D Durbin, Ryan D Wilson, Joseph M Norbeck, J. Wayne Miller, Tao Huai, Sam H Rhee: Estimates of the Emission Rates of Ammonia from Light-Duty Vehicles Using Standard Chassis Dynamometer Test Cycles. Atmospheric Environment Vol. 36, Issue 9, March 2002, pp. 1475-1482

[15] Alfréd Csikós, István Varga: Real-Time Estimation of Emissions Emerging from Motorways Based on Macroscopic Traffic Data. Acta Polytechnica Hungarica 8:(6) pp. 95-110 (2011)

Who Does Generate e-WOM and Why? – A Research Proposal

Melinda Majláth

Institute of Economics and Social Sciences, Óbuda University
Tavaszmező utca 17, H-1084 Budapest, Hungary
majlath.melinda@kgk.uni-obuda.hu

Abstract: E-WOM is a very popular topic among marketing researchers as it gives the chance for consumers to share their reviews almost totally freely. As more and more consumers' purchase decisions rely on the experiences of others shared on the Web, it has become more important to know what the motivation pattern behind the review writing activity is. There are three research questions in the focus of the present research proposal: (1) What kinds of personality traits are typical for those who generate e-WOM? (2) Why e-WOM is generated? (3) Can personality traits forecast with higher probability whether positive or negative reviews will be posted? The conceptual framework has been developed on an interdisciplinary basis: it is a combination of differential psychology and marketing. The big five personality traits (neuroticism, extraversion, agreeableness, conscientiousness, and openness to experience) are included in order to explain the intention to write electronic reviews. Moreover, a new variable called perceived informational effectiveness is introduced as a potential predictor of e-WOM generation activity.

Keywords: e-WOM; personality traits; perceived informational effectiveness; NEO-FFI

1 Introduction

In the WEB 2.0 world, word of mouth has been playing a more important role than ever before in influencing consumer decisions. The opinions of others, especially the opinions of reference persons and groups, have always played a significant role in these decisions, but now it is possible to know the opinion of hundreds and thousands of other, so-called everyday, people, people who consumers do not know personally or at all, who may well live on the other side of the world. Taking into account that cost of disseminating information online is regularly lower than offline, it is not surprising that people are ready to communicate and share information with others using the Web.

Therefore it is not surprising that marketing managers are interested in the nature of e-WOM from many different perspectives. One of these aspects is the question of how strong this informational source is, compared to the strength of information received via conventional communication channels [1] [2]. Another potential aspect is whether there is a trade-off between these channels [3] or whether they can improve each other's efficacy [4]. The third group of questions occurs in connection with the valence of e-WOM: whether negative and positive e-WOM spread over the same or different paths, whether there are different life spans, and whether there are different influential effects [1]?

From the e-WOM generation point of view, the focus is on the motivation of customers to share their opinions on WEB. They can share their experiences with people who they know, e.g. on social network sites or via e-mail, and with people who they don't know at all, e.g. on brand sites as comments, in chat forums or on a personal blog. Experts [5] investigated the forwarding motivation of internet users, with a special focus on personality traits and on the amount of internet consumption. Moreover, product attributes (such as originality and usefulness [1]) have been examined from the point of view of the question of which attributes are more likely to generate e-WOM.

This article concentrates on the formation of e-WOM, especially from the perspective of the personality of the person who writes review on the Web and, not unimportantly, on the type of the reviews they write.

2 Literature Review

According to Henning-Thurau et al. (2004), the online word of mouth is "any positive or negative statement made by potential, actual, or former customers about a product or a company, which is made available to a multitude of people and institutions via the Internet" [6] Although services are not mentioned here, we can use this definition for service evaluations as well.

The two, easily measurable features of word of mouth communication, as determined by Harrison-Walker, are its amount (how many people how intensively speak about the product or company) and its valence (whether the message, the opinion is positive or negative, and how strong it is) [7].

Moldovan et al. [1] highlighted in their research that product originality, its uniqueness, is responsible for the amount of WOM, which can be either positive or negative, while its usefulness primarily determines the valence of WOM, though it can improve the amount of WOM communication as well. Therefore, they hypothesised that the higher the usefulness, the more positive the WOM is, and they found support for this relationship in their study that examined twenty new products including electronics, hedonic instruments and furniture. Another finding was that the more original the product, the higher the amount of WOM; however, it has no significant effect on the valence of WOM. The authors argue that "WOM is spread about original products because they are interesting and about useful products because they are important" [p. 116].

I would like to add that when consumer expectations are high due to active marketing communication that emphasizes the positive or outstanding features, i.e. the product or service's uniqueness, and then it cannot satisfy these high expectations in practice, this dissatisfaction and disappointment in turn will probably generate intensive WOM.

From the receiver side there are other important aspects of e-WOM. In a laboratory experiment Gupta and Harris (2010) attempted to analyze how the presence and the increasing amount of e-WOM recommendations combined with the level of motivation for information processing can influence product choice. Different methods of information processing can influence how people view e-WOM. According to Areni et al. (2000), "The amount of thought can range from diligent consideration of topic relevant information (the central/systematic route of persuasion) to the less cognitively taxing method of association of the focal object with some positive or negative peripheral/heuristic cue" [8, p. 1043].

Those respondents who showed a higher need for cognition spent more time on considering the different products; moreover, the more e-WOM available, the more time was spent on the consideration of the alternatives (they manipulated the amount of e-WOM at three levels: none, one or ten). In contrast, when need for cognition is lower for respondents, they can rely more on e-WOM than factual information; therefore they are ready to make suboptimal product choices. Moreover, e-WOM proved to be able to shift product choice from stated preference to another product attribute-level (in the study laptop screen size was manipulated). Interestingly, a change in preference was also experienced among those who had high motivation for information processing [8].

What are the features of a useful e-WOM? Wei and Watts (2008) found that review quality and perceived source credibility are the most important motivations for adapting to online information. [9] Racherla and Friske (2012) tried to answer this question by analyzing those who use these e-WOMs for decision making [2]. Not surprisingly, they determined that the features of intangibility, heterogeneity, perishability and inseparability increases the need for additional, and primarily experimental, information to ease decision making and/or reliance on online reviews to assess services prior to use [10]. The three types of services, with different level of perceived risk and uncertainty, analyzed in the study are: search, experience and credence services, based on the typology of Darby and Karni [11]. The authors differentiated message factors from source factors, and the latter is described by identity disclosure, expertise and reputation. The factors of the message investigated are: elaborateness and review valence [2]. After analyzing 3,000 reviews from Yelp.com, they found that the reviewers' reputation and expertise positively correlated with the perceived usefulness of the review. An interesting finding was that very negative or very positive reviews proved to be more useful than others, but the length of reviews did not significantly contribute to perceived usefulness [2].

Sen and Lerman (2007) tried to explore the relationship between product type, product rating valence and perceived usefulness of the review. Based on previous studies, they supposed that readers are likely to consider negative reviews more useful than positive ones for utilitarian products (e.g. printers and digital cameras) than for hedonic products (such as music CDs and fiction books). In addition, the authors hypothesised that in the case of negative consumer e-WOM, readers will be more likely to attribute non-product related or internal motivations to the reviewer of a hedonic product than to one reviewing a utilitarian product [12].

Ahluwalia (2000) turned attention to the two-phase nature of using negative or positive reviews in decision making: first, a visitor has to decide whether he or she will pay attention to and read the review. Second, the visitor has to make a decision on whether this review will be used for the original decision making intention. [13]

To understand the motivation behind forwarding online content, Ho and Dempsey (2010) tried to explore the personal traits that may explain this behaviour, the ones that can help us here to understand better the role of e-WOM. They used Schutz's Fundamental Interpersonal Relation Orientation model (1966) as a conceptual framework. In this three-dimensional model interpersonal communication is motivated by the need to belong to a group or attract attention (called inclusion), or concern for others (called affection) or to gain power in one's social environment (called control). According to the authors, "as more people rely on the Internet as a means of communication we surmise that young adults will need to share their media experiences, particularly if they anticipate future discussions" [5, p. 1001]. Forwarding online content can strengthen the opinion leader role of a person and can help differentiate himself or herself from others. Consumers are also motivated to forward content because they care for the welfare of others: for example they want to make the others well-informed or happy. Phelps (2004) proves that an altruistic attitude can motivate online communication. [14]. Feelings of competence, achievement, influence and accomplishment can as well be generated by sharing information [15]. Therefore, those who frequently forward online content use this as "a means of developing knowledge or expertise and will be motivated by a sense of personal growth" [5, p. 1002]. The results of the structural equation modelling proved the positive relationship between individuation, altruism and the frequency of forwarding online content. However, the hypothesised positive relationship between online sharing and the need to be a part of a group was not significant. More surprisingly, personal growth/control was a significant predictor of forwarding behaviour; it was in the opposite direction. Authors explain this contrary-to-expectation outcome with the measurement type they used or with the need for feedback, which may come more directly from face-to-face communication types than from electronic ones.

The other explaining variable in the background of forwarding online content was the consumption level of electronic content, which was motivated by curiosity in the conceptual model. However, based on the results, the effect of this personal trait was not significant at all. [5]

Another question is whether the e-WOM helps to improve synergy among different communication elements of the promotion mix or whether, instead, there is a trade-off between their effects. From a general point of view increasing advertising, using mass media can improve word-of-mouth via providing topics for it. Feng and Papatla (2011) argue that increased advertising, contrary to expectations, can be associated with a reduction in online consumer word-of-mouth [3]. This negative relationship was first identified by Graham and Havlena (2007): TV adverts reduce e-WOM for soft drinks and technology, and increasing the number of online ads decreased e-WOM for the travel category [16]. The potential explanations for this phenomenon are decreasing interest caused by the more active advertising (there is no need to get more information on the product) and the customer type attracted by the given source of information [3]. The model provided by Feng and Papatla is based on the theory of Dichter (1966), who distinguished between four different types of involvement: (1) product involvement, (2) self involvement – product related conversations can satisfy emotional needs – especially when it can contribute to demonstrate superiority, or when it can give confirmation to former judgements, (3) other involvement, when the communicator tries to give something helpful to others and as a consequence, make friends, earn praise or express care or love, and (4) message involvement, which is motivated by the communication activity of producers [3]. This typology can also be useful to understand the motivation and behaviour of those who generate WOM.

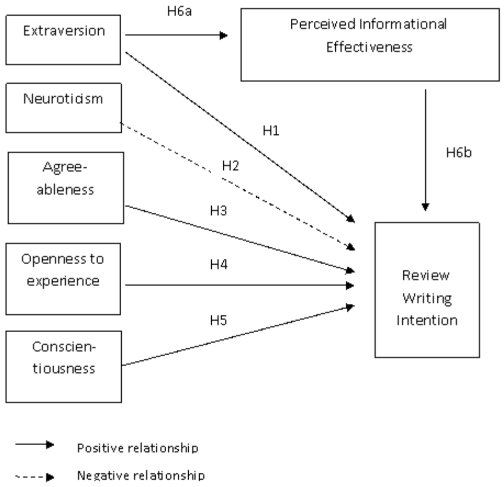


Figure 1
Conceptual Framework

3 Conceptual Framework

The results of the above mentioned studies conducted on e-WOM have raised numerous new research topics, both for companies and for consumer behaviour specialists. Here I would like to focus on the birth of these reviews.

Basically, we can differentiate three important aspects: (1) What kinds of personality traits are typical for those who generate e-WOM? (2) Why e-WOM is generated? (3) Can personality traits forecast with higher probability whether positive or negative reviews are posted?

3.1 Personality Traits

The personality traits are inevitably important antecedents of behaviour. In the differential psychological literature, experts define five dimensions of personality, called the 'Big Five', which provide a useful general framework for viewing human behaviour [17]. The elements of this construct are: (1) extraversion, (2) agreeableness, (3) conscientiousness, (4) neuroticism, and (5) openness to experience. Each domain has six facets ensuring the coverage of all aspects. The list of statements referring to neuroticism includes anxiety, hostility, depression, impulsiveness, self-consciousness and vulnerability. Warmth, gregariousness, assertiveness, activity, excitement-seeking and positive emotions are connected with the extraversion factor. Openness can be described via the terms of fantasy, esthetics, feelings, actions, ideas and values. Agreeableness consists of statements related to trust, straightforwardness, altruism, compliance, modesty and tender-mindedness. Finally, characteristics linked with conscientiousness are competence, order, dutifulness, achievement striving, self-discipline and deliberation. [18]

The significance of personal traits in connection with this topic is that opinions posted on WEB may show distortions due to the discrepancy of product users and review writers. Based on the five personality trait domains, 5 different hypotheses should be tested:

Hypothesis 1: The more extraverted the personality, (i) the more probable the e-WOM activity is, and (ii) the more positive the review is.

Hypothesis 2: The more neurotic the personality, (i) the higher the probability of the e-WOM activity is, and (ii) the more extreme the review is.

Hypothesis 3: The more agreeable the personality, (i) the higher the probability of the e-WOM activity is, and (ii) the more positive the review is.

Hypothesis 4: The more open the personality is (i) the higher the probability of the e-WOM activity, and (ii) the more positive the review is.

Hypothesis 5: The more conscientious the personality, is (i) the higher the probability of the e-WOM activity, and (ii) the less extreme the review is.

When somebody has a categorical opinion on something it is typical that he/she would like to share it if he/she extravert. Sometimes it is enough if he or she simply can add something to the ongoing discussion, giving the feeling that he/she is an “expert” in the topic, especially on forums which are organized for answering a specific question or problem from the visitors. This may serve as an explanation to our first hypothesis.

As neurotic persons can react to happenings in their surrounding impulsively, sometimes overreacting to situations, it is supposed that they will express their opinion as well via writing reviews. The sign of the review is also important – especially if there is a tendency toward posting negative reviews. On the product market, the typical sources of negative WOM are customer dissatisfaction, external media comment, competitor activity and competitive benchmarking (e.g. cost comparisons published online) [19].

As altruism was detected as a source of motivate for online communication [14], here the same direction is supposed with e-WOM activity. Conscientiousness gives the motivation to tell the truth and to share real experiences with others, and thus this trait can increase the probability of writing reviews on the Web. However, this personality trait forecasts a less volatile review pattern; they will not overreact to experiences and will probably avoid sharing exaggerated opinions with others.

3.2 Perceived Informational Effectiveness

This term is built on the construct of perceived consumer effectiveness, studied by several authors. Perceived consumer effectiveness was first examined by Kinnear, Taylor and Ahmed (1974) as the measurement of one’s belief in the results of his/her own actions [20]. The intention and behaviour of a person is the function of his/her persuasion that the occurrence of an event depends on his activity. In our context it means that an electronic review-writer believes that with sharing his or her opinion/experience, the behaviour/decision of others will change. Practically, it is the judgment of the person about the way and the extent of the impact on the environmental of his own behaviour.

Several studies [20] [21] [22] show that consumer attitudes and their reaction to messages from their surroundings are a function of their belief in their ability to positively influence the solution to the problems. Ellen et al. (1991) hypothesized that PCE represents an evaluation of the self in the context of the issue. If somebody believes that the given environmental problem can be solved by a specific action, this belief will strongly influence his/her commitment to this activity, though it can not predict other types of behaviours [23].

In this conceptual framework, it means the feeling of relativity: whether an individual can have an effect on decision of others or not; is he/or she able to influence anyone with his/her opinion or not. Built on this logic, I define perceived informational effectiveness as a belief that by expressing one's own opinion in the electronic world, one can help others to make a better decision.

Hypothesis 6a: The more extraverted the personality, the higher the perceived informational effect is.

Hypothesis 6b: The higher the perceived informational effectiveness, the higher the e-WOM generating activity is.

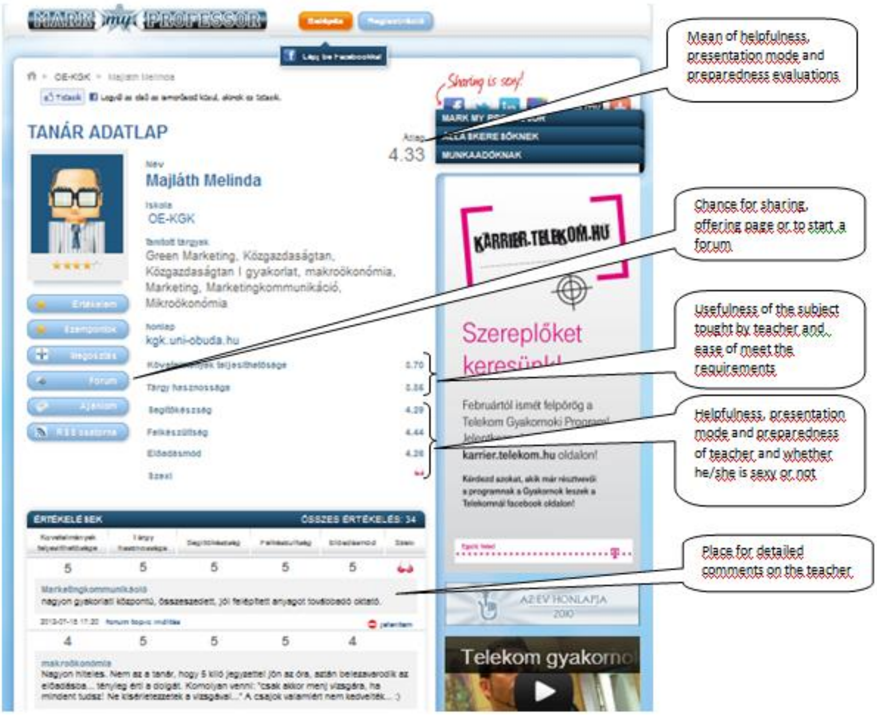
4 Research Design

As I mentioned earlier, the relevance of WOM is higher when consumers want to make a decision on services; and especially when it is an experiment or credence service that is dominated by attributes can only be evaluated after use or that be verified even then [11]. That kind of difficulty can be explained by the different tastes of individuals and intangible nature of services. Knowing this, I have chosen a topic which is completely familiar to the Y generation, as surely all of them have experience regarding teachers: I will examine the formation of e-WOM in a higher education surrounding.

The question is very up-to-date for Hungarian higher education institutions as new legislation was enacted in January, 2012. Besides other changes in structures, financing changed a lot. For example, while in the previous year the government financed the studies of 4900 students in the area of economic sciences, from September 2012 only 250 students do not have to pay a fee for their studies.[1] As more students have to contribute to the chance for taking part in higher education, they probably will be more sensitive to quality, and one aspect of it is the evaluation of teaching staff.

ISO standards and accreditation processes at higher education institutes require the evaluation of teachers by students as a part of the quality assurance. However, in addition to these official evaluations, students have other possibilities in the Web 2.0 world to comment on their teachers. One of them is markmyprofessor.com, a Hungarian site for students attending higher education institutions.

[1] http://www.felvi.hu/pub_bin/dload/FFT2012A_AOF/FFT2012A_Tajekoztatas.pdf, accessed 28.10.2012.



5 5 5 5 5

Marketingkommunikáció

nagyon gyakorlati központú, összeszedett, jól felépített anyagot továbbadó oktató.

4 5 5 5 4

makroökonómia

Nagyon hiteles. Nem az a tanár, hogy 5 kiló jegyzettel jön az óra, aztán belezavarodik az előadásba... tényleg érti a dolgát. Komolyan venni: "csak akkor menj vizsgára, ha mindent tudsz! Ne kísérletezzetek a vizsgával..." A csajok valamiért nem kedvelték... :)

Mean of helpfulness, presentation mode and preparedness evaluations

Chance for sharing, offering page or to start a forum

Usefulness of the subject tought by teacher and ease of meet the requirements

Helpfulness, presentation mode and preparedness of teacher and whether he/she is sexy or not

Place for detailed comments on the teacher

Figure 2
Structure of markmyprofessor.com page

Based on a Sen and Lerman study [12], attending courses shows utilitarian rather than hedonic features. Although they analyzed products, the importance of e-WOM for services as source of a-piori information should be more useful. Therefore, it is expected that readers will find negative reviews more helpful than positive ones. Moreover, Hungarians are said to be a pessimistic nation, who likes complaining all the time, so it is expected that more negative reviews will be provided by respondents than positive ones altogether.

In connection with teachers, below expected lecturing performance and exam difficulty are the most typical reasons for generating negative WOM. Positive features might be the teachers' style, his or her helpful behaviour and easy exams. We have to emphasise that the editors of the page have the right to revise opinions if they are scurrilous.

4.1 Research Sample

Sometimes the usefulness of academic research articles is criticized on the student sample they use; however this is not the case when the research topic itself related to Internet usage. There is a generation now that has a special attitude towards the Internet: they are called the Y generation and they form the majority of students in higher education institutes.

Members of Y generation (also known as Millennials) were born between 1980 and 1995[2]. According to Tari (2010), they are the children of the digital world. They get used to using different media at one time: they use the computer, a mobile phone, and the Internet on a daily basis, and it is almost impossible for them to manage their lives without these gadgets. They have to share their attention all the time and media addiction can be experienced by a not negligible proportion. They are open to technological innovations. Using social-networks (the most typical ones in Hungary are iwiw and Facebook) is the part of their everyday life. Writing blogs and sharing content is a form of social life. [24]

Therefore, it is not surprising that undergraduate and/or graduate student samples have been used in studies examining e-WOM [5], [8], [1], [12]. I also plan to ask undergraduate students from one Hungarian university. The planned sample size is 400 respondents. In order to gain comparable answers, I will choose students attending the same lessons in a given semester with the same teachers, and their providing different opinions on teachers despite the same experiences can be traced back to different personalities and, of course, to different expectations.

However, to get reliable and true answers, anonymity is crucial when students have to evaluate their teachers.

4.2 Variable Measurements

4.2.1. Personality Traits

As was mentioned earlier, the five personality traits which will be used as an antecedent of e-WOM generating behaviour are extraversion (E), agreeableness (A), conscientiousness (C), neuroticism (N) and openness to experience (O).

There are different operationalisations of their measurement. The most frequently studied version is the revised NEO Personality Inventory created by Costa and McCrae (1992). However, it comprises 240 statements, which would be really exhausting to evaluate by a respondent. Therefore, there have been numerous attempts to shorten the revised list further, while keeping its positive features and reliability. The NEO Five Factor Inventory [17] includes 60 items (12 items for each domain). [18] In this construct, there are different statements such as: ‘Most people I know like me’ (A), ‘I believe we should look to our religious authorities

[2] Generation Z (also known as iGeneration or the Internet Generation) will take part in higher education and the labour market in the near future, and their attitude toward eWOM should be investigated in the next phase of this research.

for decisions on moral issues' (O) or 'I often get angry at the way people treat me' (N), and the level of agreement with these statements is evaluated on a 5-point scale, where 1 means strongly disagree and 5 means strongly agree.

Therefore, for measuring personality traits, I plan to use the 60-item list of the NEO-FFI.

4.2.2. Perceived Informational Effectiveness

As I mentioned earlier, this term refers to the belief that with expressing one's own opinion one can help others to make a better decision.

In general, perceived consumer effectiveness, which is used here as a 'benchmark', was measured by statements on Likert scales. PCE is typically measured as a general concept, not in connection with specific environmental problems and their solutions. Ellen et al. used a two-statement list: 1) There is not much that any one individual can do about the environment, (2) The conservation efforts of one person are useless as long as other people refuse to conserve [23]. Berger and Corbin used 3 items: (1) I feel personally helpless to have much of an impact on a problem as large as the environment, (2) I don't feel I have enough knowledge to make well-informed decisions on environmental issues, (3) I expect the environment to continue to deteriorate until it is almost unlivable before enough attention is paid to improve it. [21]

Taking into account the proven reliabilities of these scales, it is useful to use these statements for testing our hypothesis, of course, reformulating them to the given topic. Therefore, perceived informational effectiveness will be measured by the following statements: (1) There is not much that any one reviewer can do to inform others using the Internet (2) The review of one person is useless as long as other people don't post reviews on the same topic.[3] The level of agreement with these statements will be measured on a 7-point Likert scale. Keeping it in mind that these statements are negatively formulated ones, we will have to convert some results.

4.2.3. Dependent Variable: Review Writing Intention

I intend to measure review writing intention in two ways. The first way examines review writing activity in the past, and as a consequence, it gives a broader view on respondents' behaviour. The respondent will be asked to answer the following questions: (1) How often per month do you write reviews on the internet on products or services? (2) Taking into account your reviews written in the last year,

[3] One of the referees of this article suggested to use more than two statements for measuring perceived informational effectiveness, but it would increase artificially the Cronbach alpha of the measurement and therefore would result in a bias.

what proportion (%) of your reviews were negative? (3) What was the topic of the last review you wrote? (4) Was it negative or positive?

The second method of testing review writing intention in this study will be a simulation: the markmyprofessor.com page will be introduced to the respondent and (s)he will be asked whether (s)he wants to write a review on this page or not about her/his teachers in this semester. If the answer is yes, we ask the respondents to create this/these review(s). We provide as much time as needed for each respondent.

4.2.4. Review Writing Motivations - Perceived Intention of Others

Sometimes it is more appropriate to ask people indirectly about the research topic, especially when the topic is uncomfortable, intimate or burdened with lots of social expectations. A widely used technic for this is to reformulate the question in the third person because it is supposed that the answer is mainly driven by the thoughts of the respondent; eventually it is a projective technic. If the topic is not so uncomfortable or if the respondent does not feel under pressure, the difference between direct and indirect questioning remain unremarkable.

Sen and Lerman [12] measured a variable with very similar content, but those items reflected on a given review, not a general view on reviews posted on Web. Thus, to get to know the potential motivations behind review writing, two open-ended questions will be asked: (1) What do you think people share their negative experiences/opinions with others on the Web? (2) Why do people share their positive experiences/opinions with others on the Web?

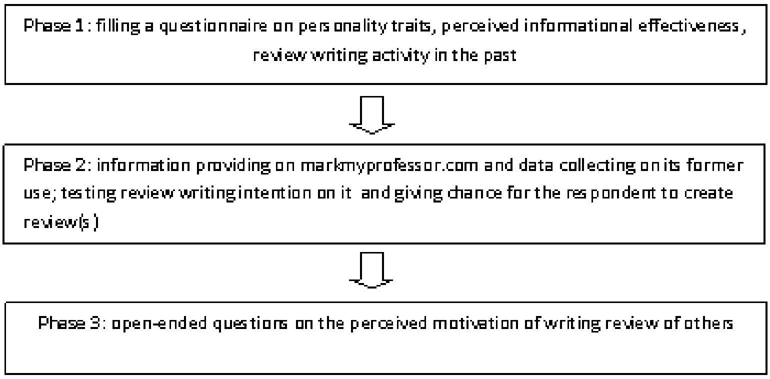

Phase 1 : filling a questionnaire on personality traits, perceived informational effectiveness, review writing activity in the past
Phase 2: information providing on markmyprofessor.com and data collecting on its former use; testing review writingintention on it and giving chance for the respondent to create review(s)
Phase 3: open-ended questions on the perceived motivation of writing review of others

Figure 3

Planned Data Collection Phases

4.3 Data Collection Phases

Data gathering will take place in three consecutive phases. First, respondents will be asked to fill out the questionnaires on personality traits and on perceived informational effectiveness and to give information on their past review writing activity. In the second phase, we collect data on the respondents' knowledge of markmyprofessor.com and ask them whether they would like to add a review on one or more of their teachers based on their experiences in the given semester. If the answer is yes, we will ask them to create this review with the numerical evaluations as well as the open-ended part (if they want to add precise details). After this phase, respondents will be asked to share their views on the motivation of others to post positive or negative reviews on the Web.

4.4 Analysis

After checking and cleaning the database, the first step will be to test the reliability of the variable constructs. Then, a factor analysis for the answers for NEO-FFI statements will be conducted, to prove the five personality trait domains; and then, based on the factor values, the H1-H5, and H6b hypotheses will be tested. Discriminant analysis will be appropriate for testing review writing intention (categorical variable). At the next level, among those who will be willing to write a review, we can use the sign (valence) of the review as a metric variable to test the (ii) parts of H1-H5. However, to distinguish negative and positive evaluations on markmyprofessor.com, where a 5 point scale is given, we have to recode the students' evaluations: as 3 is the centre of the scale, it will be redefined as 0, 4 will be transformed to 1, 5 will become 2, and 2 and 1 will be -1 and -2. To test H2 and H5, where the extremity of these evaluations is in the focus, we will use the absolute value of the evaluation.

As we will measure the general e-review-writing activity as well, we have the chance to test H1-H5 and H6b also by regression analysis, because this dependent variable is a metric one (the frequency of writing reviews as times per month)

From our point of view, the detailed comments provided by students will not be analyzed based on their content; however a dummy variable will be defined as 0 if there are no detailed comments, and 1 if any details are given by the respondent.

Perceived Informational Effectiveness is a metric variable, and therefore H6a will be tested by regression analysis.

As the last step, we toned to code the answers for the open-ended questions on the motivation for others to write reviews on the Web. And after that, by creating cross tables we can check whether there is a significant difference based on the dichotomous variable (writes a review/does not write a review on teacher(s)).

Conclusions

The future results of this planned study can help researchers to understand the attempts to generate e-WOM. Its usefulness for practitioners can be to explore the background of potential distortion in opinions shared by the internet community.

The results may also help higher education institutes to manage positive and negative WOM, as it may become a very important informational source for choosing courses and universities in the near future. Although Williams and Buttle (2011) suggested conducting systematic WOM management, it gives the misleading feeling of controllability and, to some extent, drivability of WOM. [19] However, its very basic nature is that it cannot be controlled, only sometimes, partly at the starting point.

Including commenting habits in general into this study can help to understand more precisely why people are motivated to start e-WOM. Four aspects of commenting habits will be measured: frequency, on what topics and why they write reviews and what is the typical tone of it: negative or positive. However, it would be also useful to know what the after life of these comments can be: is it important for the reviewer to check how many people found his/her comment useful or what kind of reactions is generated from others. Unfortunately, the markmyprofessor.com page, which we are using for the research, gives no opportunity to rate the usefulness of the review.

Taking this question further, another question could be whether the existing reviews influence the content of the new review the person wants to post? Do previous evaluations of the teacher influence the new evaluation? Before adding a new comment on the page, do they check the previous evaluations and do they alter the original content of the review they wanted to add?

Another further aspect could be a comparison of the evaluations of teachers in the virtual world with the official evaluations of teachers made by the university itself. Its added value would be the possibility to recognize the distorting effect of opinions posted on the web, because questionnaires for official evaluations must be filled out by the students, but on web pages only people with special personality characteristics express their own opinion, which may represent the evaluation of the minority. However, not the same aspects are measured on markmyprofessor.com and in official questionnaires, which makes this comparison impossible directly. Perhaps the ranking order based on the average evaluations may be a good basis for this task.

From a marketing management point of view, e-WOM is a part of the communication channel mix, but it has two special features: first, it cannot be controlled all the time, only partly at the beginning, when company or brand creates the message. But in this case some distrust of suspiciousness can occur, especially when something positive is being said related to the product. Second, in general it is almost costless, so it is not surprising that its effectiveness is in focus.

If e-WOM is positive, then it can work as a source to convince potential new customers or to give evidence to confirm that they made a good decision when they chose the given product. If it is negative, the uncontrollability of this stream of WOM means the biggest threat.

References

[1] Moldovan, Sarit; Jacob Goldenberg and Amitava Chattopadhyay (2011): The Different Roles of Product Originality and Usefulness in Generating Word-of-Mouth, International Journal of Research in Marketing, 28, pp. 109-119

[2] Racherla, P. and Friske, W. (2012): Perceived 'Usefulness' of Online Consumer Reviews: An Exploratory Investigation Across Three Services Categories, Electronic Commerce Research and Implications, http://dx.doi.org/10.1016/j.elerap.2012.06.003 (article in press)

[3] Feng, Jie and Papatla Pusushottam (2011): Advertising: Stimulant or Suppressant of Online Word of Mouth?, Journal of Interactive Marketing, 25, pp. 75-84

[4] Hogan, John; Katherine N Lemon and Barak Libai (2004): Quantifying the Ripple: Word of Mouth and Advertising Effectiveness, Journal of Advertising Research, 44, 3, pp. 271-280

[5] Ho, Jason. Y.C. and Melanie Dempsey (2010): Viral Marketing: Motivations to Forward Online Content, Journal of Business Research, 63, pp. 1000-1006

[6] Henning-Thurau, Thorsten; Kevin P. Gwinner, Gianfranco Walsh, and Dwayne D. Gremler (2004): Electronic Word of Mouth Via Consumer-opinion Platforms: What Motivates Consumers to Articulate Themselves on the Internet, Journal of Interactive Marketing, 18, 1, pp. 38-52

[7] Harrison-Walker, J. L. (2001): The measurement of Word of Mouth Communication and an Investigation of Service Quality and Consumer Commitment as Potential Antecedent, Journal of Service Research, 4(1), pp. 60-75

[8] Gupta, Pranjal and Judy Harris (2010): How e-WOM Recommendations Influence Products Consideration and Quality of Choice: A Motivation to Process Information Perspective, Journal of Business Research, 63, pp. 1041-1049

[9] Wei, Z. and Watts, S. (2008): Capitalizing on Content: Information Adoption in Two Online Communities, Journal of the Association for Information Systems, 9, 2, pp. 72-93

[10] Senecal, S. and Nantel, J. (2004): The Influence of Online Product Recommendations on Consumers' Online Choices, Journal of Retailing, 80, 2, pp. 159-169

[11] Darby, M. and Karni, E. (1973): Free Competition and the Optimal Amount of Fraud, The Journal of Law and Economics, 16, 1, p. 67

[12] Sen, Shahana and Dawn Lerman (2007): Why Are You Telling Me This? An Examination into Negative Consumer Reviews on the Web, Journal of Interactive Marketing, 21 (4) pp. 76-94 (published online at www. interscience.wiley.com)

[13] Ahluwalia, R. (2000): Examination of Psychological Processes Underlying Resistance to Persuasion, Journal of Consumer Research, 27, pp. 217-232

[14] Phelps, J. E.; Lewis. R, Mobilio; Perry D., Raman N. (2004): Viral Marketing or Electronic Word-of-Mouth Advertising: Examining Consumer Responses and Motivations to Pass Along Email, Journal of Advertising Research, 44(4), pp. 333-348

[15] Schutz, WC. (1966): FIRO: A Three Dimensional Theory of Interpersonal Behavior, New York: Holt, Reinehart & Winston

[16] Graham, Jeffrey and William Havlena (2007): Finding the 'Missing Link': Advertising's Impact on Word of Mouth, Web Searches and Site Visits", Journal of Advertising Research, 47, December, pp. 427-435

[17] Costa, P. T. and McCrae, R. R. (1992): Revised NEO Personality Inventory (NEO PI-R) and NEO Five Factor Inventory (NEO-FFI) professional manual. Odessa, Psychological Assessment Resources

[18] Becker, Gilbert (2005): NEO-FFI Scores in College Men and Women: A View from McDonald's Unified Treatment of Test Theory, Journal of Research in Personality, 40, pp. 911-941

[19] Williams, Martin and Francis Buttle (2011): The Eight Pillars of WOM Management: Lessons from a Multiple Case Study, Australasian Marketing Journal, 19, pp. 85-92

[20] Kinnear, T, Taylor, J. R.; Ahmed, S. (1974): Ecologically Concerned Consumers: Who Are They? - Journal of Marketing, 38 (April), 20-24

[21] Berger, I. E., Corbin, R. M. (1992): Perceived Consumer Effectiveness and Faith in Others as Moderators of Environmentally Responsible Behaviors - Journal of Public Policy and Marketing, 11 (2), 79-89

[22] Roberts, J. A. (1996): Green Consumers in the 1990s: Profile and Implications for Advertising - Journal of Business Research, Vol. 36, No. 3, 217-31

[23] Ellen, P. S, Weiner, J. L.; Cobb-Walgren, C. (1991): The Role of Perceived Consumer Effectiveness in Motivating Environmentally Conscious Behaviors - Journal of Public Policy and Marketing, 10 (2), 102-117

[24] Tari, Annamária (2010): Y GENERÁCIÓ, Jaffa Kiadó, Hungary 1st edition

[25] Vokounova, D. (2008): Who is The Young Generation? The Consumer Citizenship Network Proceedings from the 5th international conference "Assessing information", Tallinn, Estonia; pp. 164-172

Comparative Review of Statistical Parameters for Men's and Women's Basketball Leagues in Serbia

Laszlo Ratgeber*, Branko Markoski, Predrag Pecev**, Dejan Lacmanović**, Zdravko Ivanković****

*PTE-ETK University of Health Sciences Pécs - Doctor School, Pécs, Hungary, ratgeber@ratgeber.hu

**University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia, markoni@uns.ac.rs, pecev@tfzr.uns.ac.rs, dlacman@tfzr.uns.ac.rs, zdravko@tfzr.uns.ac.rs

Abstract: Basketball is one of the most popular sports. During a basketball game, statisticians note a large amount of information, helping coaches and players to improve their game and to analyse opponents in order to prepare for the following games. Due to amount and complexity of the information, basketball is the ideal discipline for the application of data mining techniques, especially neural networks that enable one to extract conclusions and knowledge from these data. In this paper, we compared the First senior leagues for men and for women in Serbia during the 2011/2012 season, and from these statistics, we calculated influence of certain parameters on a game's outcome. We concluded that the most influential parameter, in both leagues, is the offensive rebound. In men basketball then follows the three-point shot, the two-point shot and the one-point shot. In women's basketball, after the defensive rebound, the two-point shot follows, and then turnovers.

Keywords: data mining; neural networks; statistics; basketball

1 Introduction

Basketball, as a team sport, puts specific demands on players at particular positions, regarding their anthropological status. Therefore, the players' selection according to certain criteria is one of main functions of the coach. Playing experience, arising from basketball practice, points that every position in the game demands a certain level of development for certain dimensions of the players' anthropological status, which has an influence on the efficiency of playing basketball. At the beginning of a season, coaches are mostly interested in using different statistical reports for analysing and evaluating individual players. Once they have insight into the strengths and weaknesses of their players, their interest then moves towards the team as a whole. They want to know how good the team is. The team statistics therefore become most important. After all, a basketball is still a team sport. Finally, bearing in mind that different statistic reports may be used also to analyse the opponent's play, during the season the coaches' interest moves towards their opponents. More often than not, analysing the opponents' play well means the difference between winning and losing. Aspiration to win causes insight: that every detail of the game enables way to the coveted triumph. Information hunger slowly rises regarding the play and players of opponent teams, in order to better prepare for the following game, and this enables their team to impose their style and tempo in order to win. By gathering information regarding the opponent's players and their game, we begin to build a game philosophy, system and technology of scouting the opponent. By gathering information, the analysis and the systematization, we aspire toward a working model that will include, apart from gathering information, their presentation to the team, practicing individual and team tactics during micro-cycles between games and the control of its success and application during the game itself. Coaches are not the only ones to use statistics. On the contrary, whole population of sport fans and audience is able, using mass production of technology and media as TV and Internet, to follow the efficiency of teams and individual athletes. To many people, such as reporters and commentators, statistics is an important tool in doing their job, and for some, such as sport managers, it is a vital part of their profession. After all, over the last few decades, sport has become more than a game; it is a large business, with considerable amounts of money invested [3]. Professional sports organizations are multi-million-dollar companies and certain decisions are worth large amounts of money. With this kind of capital, a single wrong decision may potentially set them back years. Due to high risk and need to make correct decisions, the sport industry is just the right environment to apply data mining technologies. The final result is not the only thing important anymore.

2 Data Analysis in Sports

Huge amounts of data are present in every sport. It is extremely important to determine which information to store and to find a way for its best usage [1]. By finding the appropriate methodology to extract sense from information and to turn this information into practical knowledge, sports organizations provide themselves an advantage in comparison to other teams [2]. Such approaches to knowledge seeking may be applied into a whole organization – from players who may improve their performance using techniques of video analysis, to scouts who use statistical analysis and projection techniques in order to identify which talent will develop the most and become a good player. Most sports organizations use a third or fourth type of connection between data and their use, while only a handful uses data mining techniques. Although introduction of data mining in sports is relatively recent, the influence of those teams who applied these techniques is extraordinary [4]. Evaluations are being done based on strong analyses and scientific investigations. Since more and more sports organizations have embraced the digital era, it is possible that sports will soon become a struggle for better algorithms or better metrics for performances measurement, and analysts will become equally as important as players. The essence is in finding knowledge included in given data [5]. Statistics may also sway decisions in the wrong direction, if there is no knowledge regarding the basis of a problem, as a consequence of imprecise measurements of performances or due to an over-enhancement of certain qualities by the sports community [6]. For example, a certain player may have exceptional individual statistics, but he still may have little influence on the team as a whole. Sports statistics suffers from imprecision, since statistical metrics may not measure completely the influence of all players. For instance, a defensive rebound is a measure how many times a certain player in defence catches a ball after an unsuccessful shot by the opponent players. In order to have a defensive rebound, another player from his team must block the opponent players and therefore they are equally important in this action. Having in mind the way of noting rebounds, only the player who caught the ball is noted in statistics and rewarded a defensive rebound. The practice of data mining application in organized sports was not developed overnight. There were certain events over the past several decades, slowly bringing these changes. During 1980s, Dean Oliver began to ask questions regarding basketball metrics. His idea was to create statistics for a whole team, and not just for individual players. Oliver published his thoughts and created performances measurements for the rest of the basketball community [7]. In his work, he was especially interested in the performance of players within a team and the performance of the team as a whole, and also in how certain players function together. His work was recognized by professional sports organizations and he was hired as a consultant by the NBA team, the Seattle Supersonics. Sport analysis is usually connected to the moves of a player on the court. Following objects and trajectory analysis usually start from the analysis of the video from a game, where reference points are recognized (the size of a player is significant in comparison to that of the ball so they are easily distinguished) [8] [9]. Exposing hoaxes in sports is also possible using data mining. The NCAA organization revealed that 1% of all games played are fixed [10]. Exposing hoaxes is a relatively complex operation with data [11]. When basketball game data are collected, the first step is to find knowledge in this information. The creating of predictions has been the goal for numerous individuals and organizations for a long time. Making predictions is comprised of a number of techniques, but simulation and machine learning are the core techniques [12]. Neural networks are the most dominant system in machine learning used in sports [3]. Data sets collected from games are analysed using neural networks in order to find patterns and tendencies due to competition and financial gain. Other techniques are decision trees, genetic algorithms, ID3, and a regressive variant of the Support Vector Machine (SVM) called the Support Vector Regression (SVR).

3 Data Mining

Large quantities of data exist in all areas of sport. These data may show the individual quality of a player, the events that have taken place in the game and/or how the team functions as a unit. It is crucial to determine which data should be stored and how their usage could be maximized. By finding the correct way to extract the true meaning of the data in order to turn them into practical knowledge, sports organizations ensure themselves an advantage over other teams. Most of the analysed data are acquired from local database storage using data mining techniques such as C5.0 decision tree and its algorithm, while some data are acquired using Web Data Mining Methods [13] [14].

Modelling is done by using neuron nets. Neuron nets have been inspired by the recognition of the complex system of learning within the human brain, which is made of closely connected units of neurons. The incoming parameters are the following: *p1_procenat*, *p2_procenat*, *p3_procenat, def_rebound, of_rebound, assist, steal, lost and block*. The outcoming parameter is the result. Due to this fact, the net has nine incoming and one outcoming knots. Apart from this, there is a hidden layer in the net. The used net is a feed-forward neuron net. Each layer of the net is connected with all the knots in the previous level as well as with all the knots in the following level of the net. The net training is done with the method of the back propagation of errors, which is based on the generalised delta rule. For every syllable brought to the net during the training, the information goes through the net in advance, so that it can anticipate the outcoming layer. This anticipation is compared with the real outcoming value of the given datum, and the difference between the real and the anticipated value is then sent back through the net in order to adjust the heavy factors and to improve the anticipation of the syllables which follow. During the net training, the incoming number of data is divided in the range 75:25, into the data which will be used throughout the net training and the data which will be used throughout testing. The point is to prevent the net from saving the input as well as to prevent the wrong result as the other data are modelled. After the net training is done with the input data, we can see how much the final outcome of the game is determined by certain parameters.

The basket consists of the hoop, which is 45.7 cm (18 inch) in diameter, set at the height of 305 cm (10 feet) and mounted on the board. A team scores points by putting the ball through the hoop during a game. Shots may be scored in several ways, and the hardest to achieve are the long-distance shots, so they bring the most points. On the floor is a line at 625 centimetres from a basket. Points (in some leagues this boundary is moved even further from the basket), and shots from outside this line bring three points. Within this line, every score brings two points.

A foul happens when a player during a game is irregularly disrupted by the opponent players. If the team committing the foul has already exceeded the limit (four fouls committed during a period or quarter) or if the foul is done during an attempt to score, then a player has a chance to score from the free-throw line. Every shot scored from this line brings one point. Players have the opportunity to try two or three free throws, depending on whether the foul was committed while the player was trying to score two or three points, respectively. When throwing free throws or when shooting for two points or three points, a player may make the shot or miss the shot, i.e. score or miss. In basketball statistics, shooting percentage is the relation between shots and scores, and there are separate percentages for one-point shots, two-point shots and three-point shots. In addition to shots, which are one of major factors in basketball, there are other important factors that decide which team the team will win. Rebounds have relatively high influence on the outcome of the game. Defensive rebounds are highly important, since they prevent the opponent from having another attack after missing the basket or from scoring easy points if they obtain the ball directly under the basket. The same goes for offensive rebounds. The assist is passing the ball to teammate who is in position immediately to score. Good assists can have effect on spectators and the team itself, enabling the team to increase a game standard. Steals and turnovers are also important, since they enable a team who intercepts a ball to score easy points from counterattack.

3.1 Data Preparation Phase

The following table presents the various basketball parameters that are very important for the outcome of a match. Column headers of Table 1 have the following meanings: field ***p1_percent*** represents shots for one point, field ***p2_percent*** represents shots for 2 points, field ***p3_percent*** represents 3-point shots, field ***def_rebound*** represents jumps on the defence, field ***of_rebound*** represents a rebound, field ***assist*** represents assists, field ***steal*** represents steals, field ***lost*** represents turnovers, field ***block*** represents blocks, and the field marked ***result*** states whether a team won or lost.

Table 1

Data obtained by Data Mining

p1_ percent	p2_ percent	p3_ percent	def_re bound	of_re bound	assist	steal	lost	block	result
0.5652	0.5789	0.3668	22	8	15	12	11	3	win
0.7273	0.3556	0.4211	20	14	9	19	12	4	loss
0.6500	0.5517	0.2083	24	15	13	5	19	1	win
0.7368	0.4722	0.3333	20	6	14	6	25	6	loss
0.8421	0.6061	0.5333	13	4	11	8	13	2	loss
0.6538	0.4773	0.2857	14	8	12	9	12	2	win
0.8077	0.4958	0.3847	24	12	8	13	18	1	loss
...	...	...	...	...	...	...	...	...	...

This table shows that the values for offensive and defensive rebounds, assists, turnovers, steals and blocks are whole number values in a certain interval. The last column shows the result, i.e. the final score of the game, telling whether a team won or lost the game.

3.1.1 Effect of Certain Parameters on Outcome of the Game

The following graphs were obtained by analysis of statistical parameters in the Serbian First basketball league for men and the Serbian First "A" Basketball league for women. The First basketball league for men is a professional senior league, with 120 games played in season 2011/2012, including play-off and play-out. The First "A" Basketball league for women is a professional senior league for women, with 110 games played in season 2011/2012.

On the acquired statistical data, the analysis was done, showing influence of particular parameters on game outcome. The same parameters were observed in both leagues, with aim of doing comparative analysis. In all graphs, the colour blue represents the number of losses and red is the number of wins gained by teams in the period observed.

Effect of the One-Point Shot

Figure 1 below shows the effect of the one-point shot on the final outcome of the game. The upper graph is for results acquired for the Serbian First basketball league for men, and one below is for First "A" basketball league for women. The upper graph shows that the number of wins rises sharply when one-point shot percent exceeds 65% limit. If one-point shot percent is below 65%, teams in most cases lose that game. From the dependencies obtained, we may say that one-point shot has a significant effect on final outcome of the game.

This graph also shows that during a season, sometimes relations between shot percent and a game outcome deviates from the general rule. It happened once that a team had one-point shot percent over 85%, and they still lost the game. There was also a game when a team had one-point shot percent of only 38%, and still they won. Statistical minimum for one-point shot percent is 25%, and statistical maximum is 96.2%. The average value for this shot, regarding all games in the league, was 71.4%. The standard deviation is 0.11.

In the Serbian First "A" basketball league for women, the one-point shot percent does not influence final score of the game so much. This is visible from the lower graph, since an increase in shot percentage with the number of wins rises slower. In addition, with a decrease in the percent, the number of losses gradually increases.

The statistical minimum for one-point percent is 0%, which means that a team did not score even one free throw, and the statistical maximum is 95.2%. The average value for this shot, regarding all games in the League, is 66.8%. The standard deviation is 0.125.

For every coach, the one-point shot is very important. At every break during games, players are prompted to practice free throws, in order to improve this segment of their game.

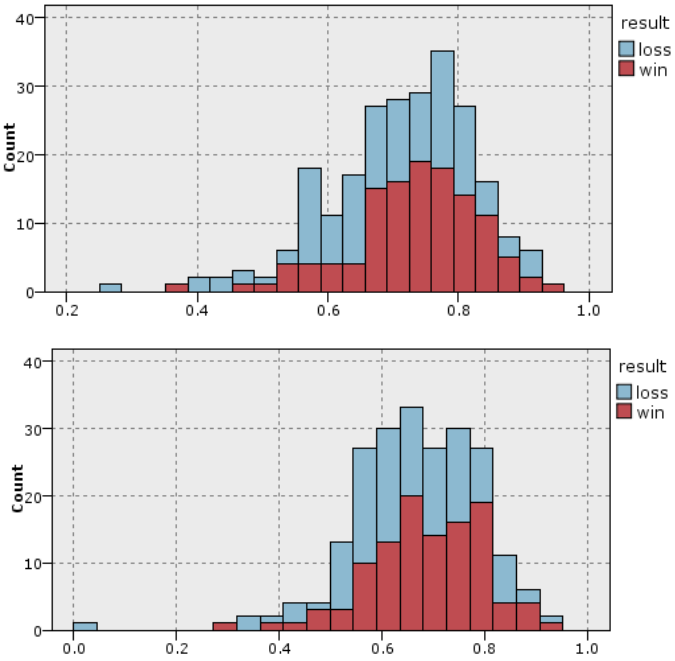


Figure 1

Effect of one-point shot percent on the final outcome of the game in Serbian First basketball league for men (above) and the Serbian First "A" basketball league for women (below)

Effect of the Two-Point Shot Percent

Figure 2 shows effect of the two-point shot on the final outcome of the game in the Serbian First basketball league for men (above) and the Serbian First "A" basketball league for women (below). It is visible that number of wins is higher when the two-point shot percent is over 58%. Below this level, the number of losses is higher. Therefore, the two-point shot percent has an effect on the game outcome. Here also are exceptions from general rule. It happened that teams had a two-point shot percent over 70%, and they still lose the game. In addition, it happened that a team with only 40% of two-point shots won the game.

The statistical minimum for the two-point shot percent 39.5% and statistical maximum is 83%. The average value for this shot, regarding all games in the league, is 58.2%. The standard deviation is 0.083. The two-point shot percent has a higher minimal value and a lower maximal value than the one-point shot.

Regarding women's league graph, it is visible that number of wins is higher when the two-point shot percent is over 50%. When the two-point shot percent is below

40%, the number of losses is considerably higher. Therefore, this percent has a significant effect on the game outcome. By comparing these graphs, we may conclude that the two-point shot has more influence in women's basketball than in men's.

These graphs also show particular values that do not comply with the general rule. It happened that a team had a two-point shot percent over 83% and still lost the game. It has also happened that a team with about 30% still won the game.

The statistical minimum for the two-point shot is 24.3%, and the statistical maximum is 85.2%. The average value for this shot, regarding all games in the league, is 48.3%. The standard deviation is 0.105.

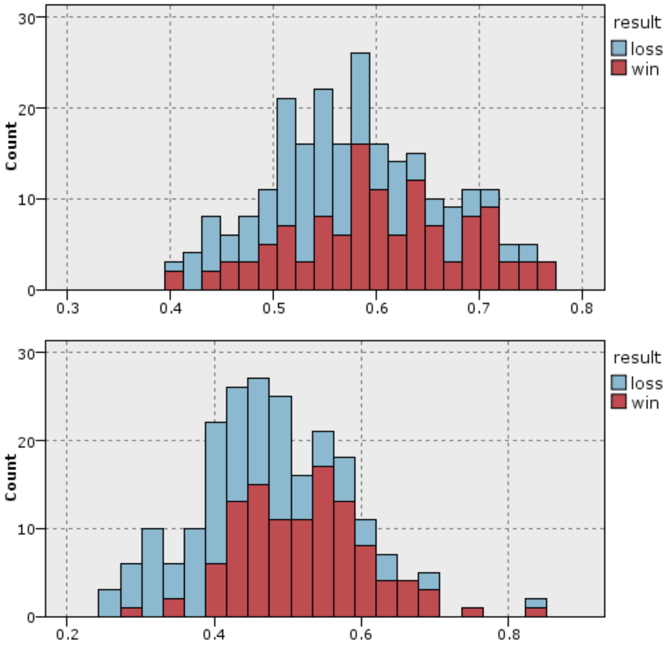


Figure 2

Effect of two-point shot percent on the final outcome of the game in the Serbian First basketball league for men (above) and the Serbian First "A" basketball league for women (below)

Effect of the Three-Point Shot

Figure 3 shows the effect of the three-point shot on the final outcome of the game in both leagues observed. From the upper graph (the league for men), it is visible that the number of wins is higher when the three-point shot percent is over 38%. When three-points shot percent is under 30%, the team loses the game more often. From this we may conclude that three-point percent has a significant influence on the outcome of the game. The statistical minimum for the three-point shot percent is 0%, meaning that a team did not realize a single three-point shot, and the statistical maximum is 83%. The average value for this shot, regarding all games in the league, is 34.9%. The standard deviation is 0.349. The three-point shot has more extreme values than the two-point shot because of the lower numbers of shots during a game.

From the lower graph, showing the three-point shot percentages for the Serbian First "A" basketball league for women, it is visible that the number of wins is higher when the shot percent is over 43%. If the percent is over 60%, the team wins all games. When the shot percent is under 43%, there is less influence on the game outcome, since the number of wins and losses in these situations is approximately the same. From this we may conclude that the three-point shot percent has an effect on the outcome of the game, but this effect is significantly lower than in the Serbian First basketball league for men.

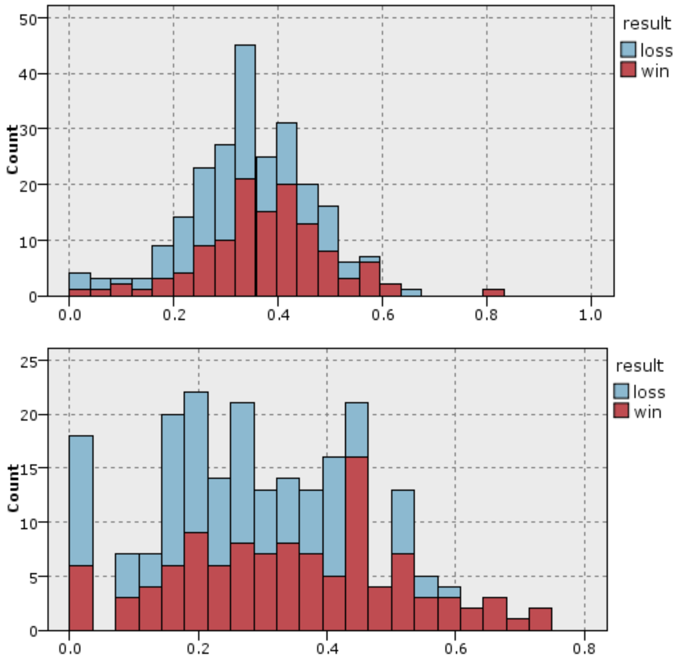


Figure 3

Effect of the three-point shot percent on the final outcome of the game in the Serbian First basketball league for men (above) and the Serbian First "A" basketball league for women (below)

In the sample observed, there was the extreme case when a team had a three-point shot percent of 0%, without a single successful shot, and still they won. This happened in 6 out of 110 observed games.

The statistical minimum for the three-point shot percent is 0%, and statistical maximum is 75%. The average value for this shot, regarding all games in the league, is 30%. The standard deviation is 0.166.

Effect of Defence Rebounds

There is a dilemma in some situations: determining whether it is a rebound or a steal. There are obviously offensive and defensive rebounds, but what about a situation when no one gets the ball in rebound? Many coaches think that if a ball is "rolling" on the court after being deflected from the hoop, a player who takes it must have a steal and not a rebound. Some think that steal should be given to the player who deflects a ball to his teammate, if he is not able to catch it after shot. If this is steal, then is the so-called tap in an offensive rebound or a steal?

Figure 4 shows the effect of defensive rebounds on the win or loss of the team. The graph for the Serbian First basketball league for men shows that the number of wins is higher when the number of defensive rebounds is over 22 for the team.

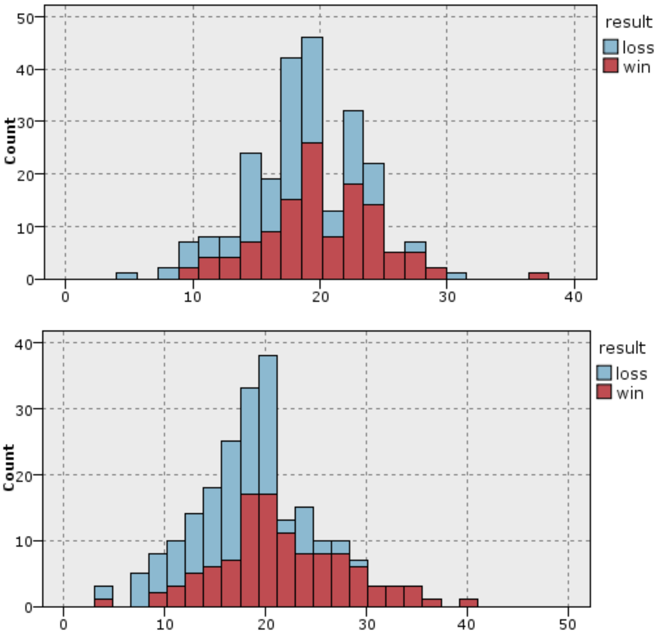


Figure 4
Effects of defensive rebounds on the final outcome of the game in the Serbian First basketball league for men (above) and the Serbian First "A" basketball league for women (below)

If the number of defensive rebounds is under 10, the team will lose the game. Regarding this distribution of wins and losses depending on defensive rebounds, we may conclude that they have a significant effect on the game outcome. With an increase in rebounds, the number of wins also increases, and vice versa.

The graph shows also a case when the team had 31 defensive rebounds, and they still lost the game. The statistical minimum for defensive rebounds is 4, and the maximum is 38. The average value, regarding all games in the league, is 18.98. The standard deviation is 4.637.

From the graph for the Serbian First “A” basketball league for women it is visible that the number of wins increases when the number of defensive rebounds is over 22 for the team. If the number of defensive rebounds is under 16, the team will most often lose the game. Therefore, we may conclude that defensive rebounds have a significant effect on the outcome of the game, similar as in the Serbian First basketball league for men. The statistical minimum for defensive rebounds is 3, and the maximum is 41. The average value regarding all games in the league is 19.23. The standard deviation is 6.213.

Effect of Offensive Rebounds

Figure 5 shows the effect of offensive rebounds on the final outcome of the game. From the upper graph (the Serbian First basketball league for men), it is visible that the number of wins is higher when the number of offensive rebounds is over 15. Because of the almost identical number of wins and losses, offensive rebounds do not have much effect on the final outcome of the game. There is also the case when the team had 24 offensive rebounds and they still lost the game.

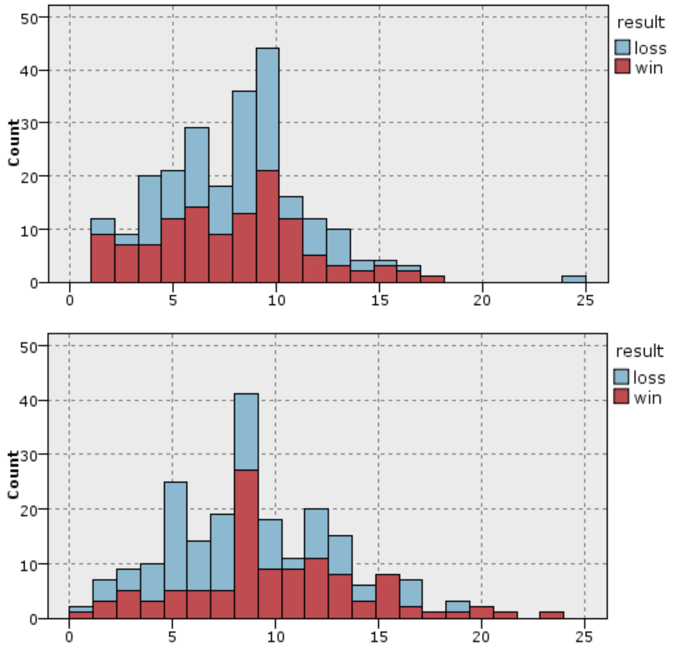


Figure 5

Effects of offensive rebounds on the final outcome of the game in the Serbian First basketball league for men (above) and the Serbian First “A” basketball league for women (below)

The statistical minimum for offensive rebounds is 1, and the maximum is 24. The average value, regarding all games in the league, is 7.93. The standard deviation is 3.529.

From the graph for women's basketball (below) it is visible that a team will certainly win if they have over 20 rebounds. Offensive rebounds have no major effect on winning or losing in this league either. There was a case when a team had 18 offensive rebounds, and still they lost the game. There was a game when one team had no offensive rebounds, and yet they won the game. The statistical minimum for offensive rebounds is 0, and the maximum is 24. The average value, regarding all games in the league, is 9.04. The standard deviation is 4.264.

Effect of Assists

Figure 6 shows the effect of assists on winning or losing by a team in the Serbian First basketball league for men (above) and the Serbian First "A" basketball league for women (below). In the men's league, when there are over 15 assists in a game, the team is wins in most cases. When a team has less than 8 assists, they will mostly lose.

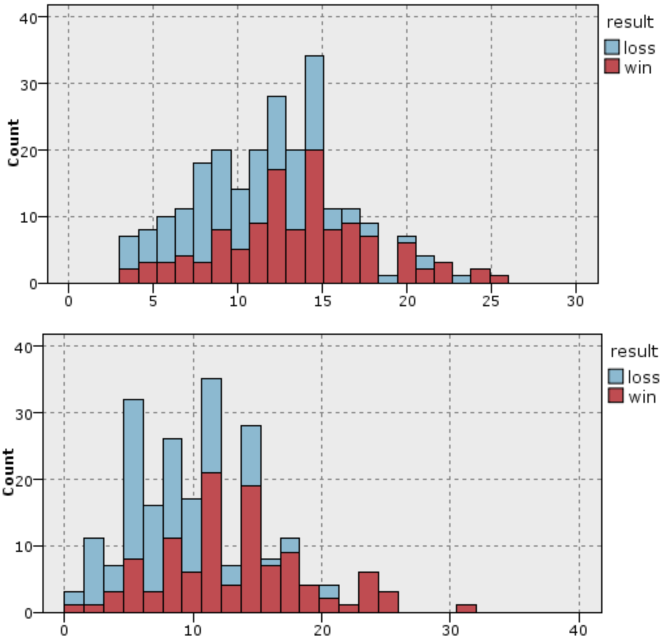


Figure 6

Effects of assists on the final outcome of the game in the Serbian First basketball league for men (above) and the Serbian First "A" basketball league for women (below)

The graph also shows cases when a team had 19, 20, 21 or 23 assists, and yet they lost the game. We may conclude that assists have no major effect on the outcome of the game. The statistical minimum for assists is 3, and the maximum is 26. The average value, regarding all games in the league, is 12.08. The standard deviation is 4.472.

In the women's league, when the number of assists is over 11, the team will mostly win. If a team has fewer than 8 assists, they will lose the game in most cases.

The graph also shows that a team who has more than 22 assists certainly wins the game. There was the case when a team had no assists, and they still won the game. The number of assists has more effect in the women's league than in the men's.

The statistical minimum for assists in the women's league is 0, and the maximum 32. The average value, regarding all games in the league, is 10.87. The standard deviation is 5.537.

Effect of Steals

Figure 7 shows the effect of steals on wins and losses in games in the two leagues observed. When the number of steals is over 15 in the men's league, this team will mostly win. If the number of steals is under 6, this team will mostly lose the game. From this graph, we may conclude that steals have an effect on outcome of the game, but less than defensive rebounds or shot percentages.

The statistical minimum for steals is 0, when a team has no steals during a game, and the maximum is 21. The average value regarding all games in the league is 9.18. The standard deviation is 4.192.

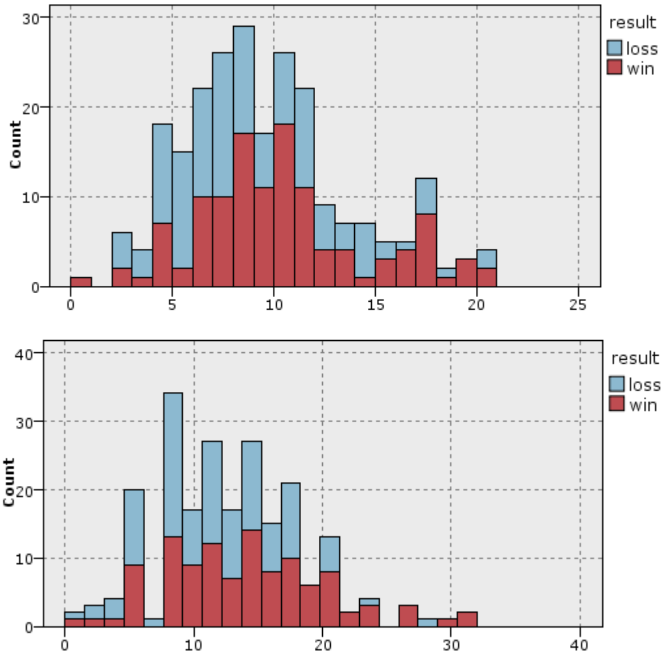


Figure 7
Effect of steals on the final outcome of the game in the Serbian First basketball league for men (above) and the Serbian First "A" basketball league for women (below) steals

In women's competition, when the number of steals is over 18, the team will mostly win. With a decrease in the number of steals, the number of losses increases.

The statistical minimum for steals is 0, and the maximum is 32. The average value regarding all games in the league is 12.99. The standard deviation is 5.644.

Effect of Turnovers

Figure 8 shows the effect of number of turnovers in the competitions observed. In the men's league, if the number of turnovers is over 15 (graph above), the team mostly losses the game. If a team has less than 10 turnovers, they mostly will win the game. This graph shows that turnovers have a significant effect on the outcome of the game.

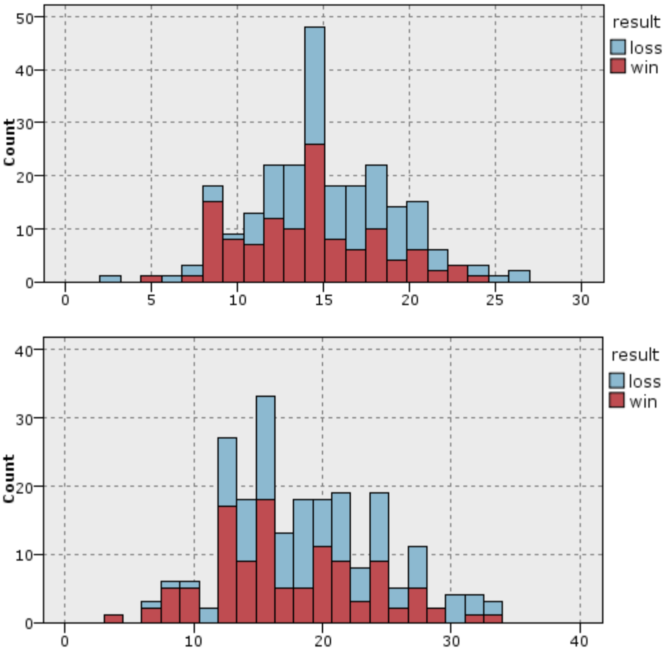


Figure 8

Effects of turnovers on the final outcome of the game in the Serbian First basketball league for men (above) and the Serbian First "A" basketball league for women (below)

The statistical minimum for turnovers is 2, and the maximum is 27. The average value, regarding all games in the league, is 14.92. The standard deviation is 4.144. In the women's league, if the number of turnovers is over 16, the team will mostly lose. If the team has less than 10 turnovers, they will mostly win the game. This graph shows that turnovers have a significant effect on the outcome of a game. The statistical minimum for turnovers is 3, and the maximum is 34. The average value, regarding all games in the league, is 18.49. The standard deviation is 6.058.

Effect of Blocks

Figure 9 shows the effect of blocks on the final outcome of the game. From the graph above, it is visible that if there is more than 4 blocks, the team will mostly win the game. This statistical parameter has no significant effect on the outcome of the game if its value is fewer than 4. The statistical minimum for blocks is 0, and the maximum is 8. The average value, regarding all games in the league, is 2.22. The standard deviation is 1.835. The graph for women’s league is almost identical to the one for the men’s league, so we may conclude that in women’s basketball blocks have no major effects on the final outcome of a game.

The statistical minimum for blocks is 0, and the maximum is 8. The average value, regarding all games in the league, is 2.02. The standard deviation is 1.728.

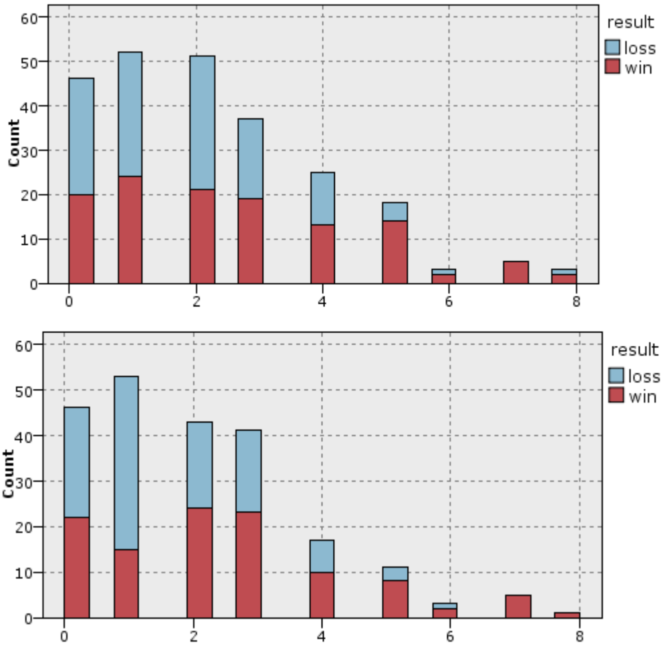


Figure 9

Effects of blocks on the final outcome of the game in the Serbian First basketball league for men (above) and the Serbian First “A” basketball league for women (below)

3.1.2 Comparative Analysis of the Basic Basketball Parameters

Table 2 shows the comparative list of one-, two- and three-point shots for the Serbian First basketball league for men and the Serbian First “A” basketball league for women. It can be seen that in the women’s league, the statistical minimum for one-point shot percent is 0%, meaning that a team did not score any free throws, and consequently they lost the game. In men’s league, this percent varies from 25% to a maximum 96.2%. Regarding the average values of shots from all positions, we can conclude that men are more precise. Precision difference varies from 4.6% for the one-point shot, to 4.9% for the three-point shot to 9.9% for the two-points shot. Considering the three-points shot, Jamie Angeli [15] says that these shots are hardest to score, since a player must be alone in order to score successfully from a distance. Many players stay after training sessions in order to perfect this shot [16] [17].

Table 2

Basic statistical parameters for one-, two- and three-point shots

	1P				2P				3P			
MEN	Min%	Max%	MEAN	S.DEV	Min%	Max%	MEAN	S.DEV	Min%	Max%	MEAN	S.DEV
	25.0	96.2	71.4	0.110	39.50	77.40	58.2	0.083	0.00	83.30	34.9	0.122
WOMEN	1P				2P				3P			
	Min%	Max%	MEAN	S.DEV	Min%	Max%	MEAN	S.DEV	Min%	Max%	MEAN	S.DEV
	0.0	95.2	66.8	0.125	24.30	85.2	48.3	0.105	0.00	75.00	30.0	0.166

Table 3 shows three very important parameters: defensive rebounds, offensive rebounds and assists. It can be seen that the average number of rebounds per game is higher for women. Women have on average 0.244 more defensive and 1.112 more offensive rebounds. This is to be expected, since they have lower shot percentages from all positions. Data regarding assists are especially interesting. According to FIBA statistician manual [18], an assists is a pass leading directly to another player’s score, and only if a shooter reacts by immediate shot movement. A pass to another player who is in a good shooting position but who considers other options before shooting, is not an assist. Shooting distance and the manner of shot are not factors deciding whether the pass is also an assist. Passing to a player who is alone on the middle-court and who dribbles to basket is assist, but if this player must dribble past an opponent player, then it is not an assist. Therefore, the deciding factor for determining an assist must be the sum of his action and of other player’s imminent intention to shoot and score. In USA, this is quite different, whether in NBA, NSA or WNBA. There the assist is noted if a ball was passed to another player when he/she is alone in a shooting position, no and it does not matter whether he/she was fouled after that. In Table 3, it is visible that men have on average 1.21 more assists per game.

Table 3

Basic statistical parameters for defensive rebounds, offensive rebounds and assists

	DEF. REBOUND				**OFF. REBOUND**				**ASSIST**			
MEN	Min	Max	MEAN	S.DEV	Min	Max	MEAN	S.DEV	Min	Max	MEAN	S.DEV
	40	38	18.988	4.637	1.00	25	7.929	3.529	3	26	12.083	4.472
WOMEN	**DEF. REBOUND**				**OFF. REBOUND**				**ASSIST**			
	Min	Max	MEAN	S.DEV	Min	Max	MEAN	S.DEV	Min	Max	MEAN	S.DEV
	3	41	19.232	6.213	0	23	9.041	4.264	0	32	10.873	5.537

The next set of parameters is a basis for good defence and a counter-attack. Table 4 shows that women had more steals than men, on average 3.816 more per game. Regarding turnovers, women are more successful too: on average, 3.57 per game. Therefore, we can conclude that in the Serbian First "A" basketball league for women, better defence is played. The number of blocks is almost identical in both leagues.

Table 4

Basic statistical parameters for steals, turnovers and blocks

	STEAL				TURNOVER				BLOCK			
MEN	Min	Max	MEAN	S.DEV	Min	Max	MEAN	S.DEV	Min	Max	MEAN	S.DEV
	0	21	9.179	4.192	2	27	14.925	4.144	0	8	2.217	1.835
WOMEN	STEAL				TURNOVER				BLOCK			
	Min	Max	MEAN	S.DEV	Min	Max	MEAN	S.DEV	Min	Max	MEAN	S.DEV
	0	32	12.995	5.664	3	34	18.495	6.058	0	8.	2.027	1.728

3.2 Model Evaluation Phase

In the Serbian First basketball league for men, the network correctly predicted 195 outcomes from 240 possible (120 games where every team must either win or lose), which is 81.25% of total number of input data. Therefore, the model failed 45 times to predict the outcome correctly, which is 18.75% of all cases. From 120 wins documented, the algorithm correctly predicted 99, while for 21 wins it predicted losses. Regarding losses, the algorithm correctly predicted 96 from 120 losses, while for 24 losses it predicted wins. The confidence matrix is shown in Table 8. In the Serbian First "A" basketball league for women, the network correctly predicted 304 from 362 outcomes or 83.98% of the total number of input data. Therefore, the model did not correctly predict 58 outcomes, which is 16.02% of all outcomes. From 181 wins documented, the algorithm correctly predicted 144, while for 37 wins it predicted losses. Regarding losses, the algorithm correctly predicted 160 losses, while for 21 losses it predicted wins. The confidence matrix is shown in Table 9.

The model including the most relevant basketball parameters has a relatively high prediction precision for game outcomes based on the input parameters. More than eighty percent of the input data would correctly predict the outcome of the game.

Table 6

Model precision for the Serbian First basketball league for men

True	195	81.25%
False	45	18.75%
Total	240	

Table 7

Model precision for the First "A" basketball league for women

True	181	82.27%
False	39	17.73%
Total	220	

Table 8

Confidence matrix for the Serbian First basketball league for men

	Loss	win
loss	96	24
win	21	99

Table 9

Confidence matrix for the Serbian First "A" basketball league for women

	loss	win
loss	82	28
win	11	99

The results obtained by the neural network were confirmed by the C5.0 decision tree.

A prediction correctness of over 80% confirms that the model used was correct. Greater correctness was prevented by the fact that, in keeping basketball statistics, a large number of events are not noted. Programs like BSV are used in real time, during the game, so there must be a selection as to which events will be documented and which not. In order to obtain more complete knowledge regarding a game and find some new patterns, we need a richer data set, or the application of software solutions that would note all relevant events during later viewing of the game.

Conclusion

The game of basketball is progressing rapidly. The number of quality players and teams is growing quickly. At high levels of competition, there are no teams that can count on a safe win for every game. Good preparation for the game may mean the difference between average and good results. Scouting opponents is an important and indispensable element in these preparations. Special attention must be paid not only to the manner of collecting and recording data during a game or by viewing video clips, but also to a way of processing data and presenting them to the team. Increasing professionalism and competition are prompting clubs to develop systematic approaches to all activities. With data mining techniques, it is possible to analyse both certain players and teams as a whole. Using modelling data from the Serbian First basketball league for men and the Serbian First "A" basketball league for women, we obtained knowledge regarding the way of playing and the decisive elements that influence the final outcome of the game. In both leagues, defensive rebounds have the most influence on winning the game. Men's basketball is based on a faster game and more shots, so precision and good training is very important. Women's basketball pays more attention to defence, so other important elements are steals and turnovers. In addition to these, it is important to maintain a high level of two-point shots, not to miss "safe shots". In team sports, there are many methods and ways of preparing athletes for competition. The results were obtained from the application of a neuron net to the collected data, whereas the check was done through applying C5.0 decision trees, which confirmed the results.

These methods include physical, technical, tactical, psychological and integral types of preparation. Each of them has fundamental importance in the formation of athletes and teams, leading to a successful performance at the competition and good results. Today, good scouting cannot be imagined without the use of modern information technologies. The coaches at the game often ask how many points were received from the zone defence and how the particular player shoots from different positions. Coaches, assistant coaches, players and scouts are also interested in how much each player scored, information about everyone's points given different types of defence, and how many points were received from another attack or counterattack, or whether the team will take on a zone defence in time. These are additional parameters that influence the final outcome of the game and which are not a subject of the modelling in this paper. The reason for this is that such information is not recorded in the kept statistics on one basketball game, but may be recorded in a deeper analysis of the game.

Acknowledgement

This work was partially supported by the Serbian Ministry of Education and Sciences (Grant No: 171039).

References

[1] Lyons, K. "Data Mining and Knowledge Discovery", Australian Sports Commission Journals 2, 2005

[2] O'Reilly, N., P. Knight "Knowledge Management Best Practices in National Sport Organizations". International Journal of Sport Management and Marketing 2(3), pp. 264-280, 2007

[3] Schumaker R., Solieman O., Chen H. "Sports Data Mining", Springer 1st edition, 2010

[4] Stefani, R. "A Taxonomy of Sports Rating Systems". IEEE Transactions on Systems, Man, and Cybernetics - Part A 29(1): 116-120. 1999

[5] Choo, C. W. "The Knowing Organization: How Organizations Use Information to Construct Meaning, Create Knowledge, and Make Decisions". International Journal of Information Management 16(5): 329-340, 1996

[6] Fieltz, L. & D. Scott. "Prediction of Physical Performance Using Data Mining". Research Quarterly for Exercise and Sport 74(1): 1-25, 2003

[7] Dean Oliver, "Basketball on paper – Rules and tools for performance analysis", Brassey's, Washington DC, 2005

[8] Chang, C. W., S. Y. Lee. "A Video Information System for Sport Motion Analysis". Journal of Visual Languages and Computing 8(3): 265-287

[9] Marakas, G. "Modern Data Warehousing, Mining, and Visualization: Core Concepts". Prentice Hall, Upper Saddle River, NJ, 2003

[10] Wolfers, J. "Point Shaving: Corruption in NCAA Basketball". AEA Papers and Proceedings 96(2): 279-283, 2006

[11] Dobra, J., T. Cargill, B. Goff, R. Tollison. "Efficient Markets for Wagers: The Case of Professional Basketball Wagering. In Sportometrics", Texas A&M University Press, College Station, TX, 215-249

[12] Witten, I. H. & Frank, E. "Data Mining: Practical machine learning tools and techniques, 2nd Edition". Morgan Kaufmann, San Francisco, 2005

[13] Shih-Yang Yang, Po-Zung Chen, Chu-Hao Sun, "Using Petri Net to Enhance Web Usage Mining", Acta Polytehnia Hungarica, Vol. 4, No. 3, 2007, pp. 113-125

[14] Lajos Izsó, Péter Tóth, "Applying Web-Mining Methods for Analysis of Student Behaviour in VLE Courses", Acta Polytehnia Hungarica, Vol. 5, No. 4, 2008, pp. 79-92

[15] Jamie Angeli, Scouting America's Top Basketball Programs, Volume 1, 2003

[16] Ratgeber, L. Play from a Game: (Head Coach). Mizo Pecs 2010, 2007/2008, Mizo Pecs 2010 vs. Euroleasing Sopron

[17] Serbian Basketball Coaches: Zeljko Obradovic, Aleksandar Dordevic, Branislav Prelevic, Bozidar Maljkovic, Dusko Vujosevic, Svetislav Pesic , Book group 2008

[18] FIBA - Basketball Statisticians' Manual, 2010

Tag and Topic Recommendation Systems

Ágnes Bogárdi-Mészöly[1,3], András Rövid[2], Hiroshi Ishikawa[3], Shohei Yokoyama[3], Zoltán Vámossy[2]

[1]Department of Automation and Applied Informatics, Budapest University of Technology and Economics, Magyar tudósok körútja 2. QB207, 1117 Budapest, Hungary, agi@aut.bme.hu

[2]John von Neumann Faculty of Informatics, Óbuda University, Bécsi út 96/B, 1034 Budapest, Hungary, rovid.andras@nik.uni-obuda.hu, vamossy.zoltan@nik.uni-obuda.hu

[3]Department of Computer Science, Shizuoka University, 3-5-1 Johoku, Naka-ku, 432-8011 Hamamatsu, Japan, ishikawa@inf.shizuoka.ac.jp, yokoyama@inf.shizuoka.ac.jp

Abstract: The spread of Web 2.0 has caused user-generated content explosion. Users can tag resources in order to describe and organize them. A tag cloud provides rough impression of relative importance of each tag within the overall cloud in order to facilitate browsing among numerous tags and resources. The size of its vocabulary may be huge, moreover, it is incomplete and inconsistent. Thus, the goal of our paper is to establish tag and topic recommendation systems. Firstly, for tag recommendation system novel algorithms have been proposed to refine vocabulary, enhance reference counts, and improve font distribution for enriched visualization. Secondly, for topic recommendation system novel algorithms have been provided to construct a special graph from tags and evaluate reference counts for topic identification. The proposed recommendation systems have been validated and verified on the tag cloud of a real-world thesis portal.

Keywords: social network; tag cloud; tag analysis; vocabulary; reference count; font distribution algorithm; topic recommendation

1 Introduction

With the appearance of Web 2.0 [1] and the spread of social media sites [2] users became from passive spectators to active content generators. Users can interact and collaborate with each other in virtual communities. Nowadays lots of social sites exist for various purposes: collaborative projects (Wikipedia), blogs (Twitter), content communities (YouTube), social networking sites (Facebook), virtual game worlds (World of Warcraft), virtual social worlds (Second Life), etc.

There are numerous books dealing with this topic denoted for introduction and marketing [3] for research [4] etc. Various significant companies have research groups for social computing: Microsoft [5], IBM [6], HP [7], etc.

Usually on these sites users can assign tags to resources in order to describe and organize them. This tagging process [8] establishes associations between tags and resources, which can be applied to navigate to resources by tags, as well as, to tags based on related tags, etc. With tagging a folksonomy (folk (people) + taxis (classification) + nomos (management)) evolves, which is the vocabulary of tags emerged by the community [9]. The size of these vocabularies may be huge, moreover, they are incomplete and inconsistent. Thus, in connection with social tagging several challenges have emerged [4].

The paper is organized as follows. Section 2 covers the background. Section 3 introduces the experimental environment. Section 4 presents the proposed tag recommendation system. Section 5 shows the provided topic recommendation system. Finally, last section reports the conclusions.

2 Background

In this section, definitions related to tagging are discussed. Tags are user-defined informal and personal strings, short descriptions related to resources, keywords associated with resources. They are helpful in browsing and searching. Resources are such identities which can be tagged, such as text, image, audio, video, document, etc. Tagging is the process of assigning existing and new tags to resources. Tag recommendation systems exist to help users in tagging based on own tags or tags of other community members.

There are lots of different kinds of tags: content-based, context-based, attribute, ownership, subjective, organizational, purpose, factual, personal, self-referential, tag bundles, etc. Furthermore, users have various motivations for tagging: future retrieval, contribution and sharing, attract attention, play and competition, self presentation (self referential tags), opinion expression, task organization, social signaling, money, technology ease, etc. [4]. In our tag clouds, tags are content-based, tagging is motivated by contribution and sharing. With content-based tags the actual content of the resources can be identified. By the contribution and sharing as motivation, tags describe resources, and add them to conceptual clusters or refined categories for known and unknown audience.

Tag clouds are visually depicted tags in order to facilitate browsing among numerous tags and resources. It gives rough impression of relative importance of each tag within the overall cloud. In this paper such tag clouds are investigated.

In some situations to answer various questions browsing in tag clouds are more useful than searching [10]. Search interface is preferred if the needed information is specific. Tag clouds are preferred if the sought information is more general.

For this reason, the visualization of tag clouds is one of the most important and complicated consideration [11]. Tag clouds have two dimensional representations. Tags can be ordered alphabetically, based on semantic similarity or any kind of clusters [12] [13]. Relevant tags can be visually emphasized using such visual properties as shape, color, position, etc. Each tag cloud is visualized in its own unique way. The basis of the used methods is similar, but there are no two tag clouds whose visualization is the same. Numerous font distribution algorithms exist [14] [15]. In our tag clouds all tags are represented simply ordered by alphabetically and visually weighted by letter size.

In social networks power laws occur many times in many contexts [4]. A random variable is distributed according to a power law when its probability density function is given by $x^{-\gamma}$, where $x \geq x_{\min}$ and $\gamma > 1$ [16]. γ is a constant parameter called exponent or scaling parameter, typically in the range of $2 < \gamma < 3$.

In our previous work, algorithms have been yielded to improve tag clouds [17]. In addition, font distribution algorithms have been investigated [18]. Moreover, algorithms have been provided to recommend topics [19]. In this paper, complete tag and topic recommendation systems are established.

3 Experimental Environment

The Faculty of Electrical Engineering and Informatics of the Budapest University of Technology and Economics has a web portal to manage all theses of the faculty for the whole workflow starting from description to review [20].

This portal has been implemented as a three-tier ASP.NET web site. The presentation layer is in HTML and jQuery. In the business logic layer there are C# classes. In the data access layer LINQ and stored procedures are used mixed. The database is in Microsoft SQL Server. The provided algorithms have been implemented in SQL stored procedures, C# classes using LINQ, and MATLAB functions [21].

On this thesis portal tags are assigned to theses to describe and organize them in order to be helpful in browsing and searching. The portal has tag clouds in Hungarian and English languages.

For fast visualization of tag clouds, the tags are stored in a cache. The cache is refreshed when a new tag is created, an existing tag is modified, deleted. The enhanced reference counts (see Section 4.2) are stored in this cache, as well.

In these clouds the tags are classified to four classes according to their reference counts. The number of classes is an arbitrary parameter, it can be varied. The originally applied font distribution algorithm uses weights based on number of classes, and average, maximum, minimum reference counts. The following proposed systems have been validated and verified on these tag clouds.

4 Tag Recommendation System

The aims of the proposed tag recommendation system are to improve the quality of tags, detect the actually popular tags, and enrich the visualization of tag clouds. The provided system has three main steps. In the first step the vocabulary is refined, namely, certain spelling and clerical errors are corrected, in addition, certain tags are contracted. In the second step the reference counts are enhanced. In the third step the font distribution algorithm is improved.

4.1 Vocabulary Refinement

In this section, two algorithms have been proposed to improve the quality of tags. The purpose of the first algorithm is not to correct all spelling and clerical errors in each tag, but to contract such tags which are the same only with different writing (the algorithm retains not in all cases the grammatically correct version). The aim of the second algorithm is to contract such simple tags of each thesis to a compound tag in which case the created compound tag is used for another thesis. Table 1 summarizes the meaning of the indices used in these algorithms. In addition, the notations of these algorithms can be seen in Table 2.

Table 1

Meaning of indices used in the algorithms of the tag recommendation system

Notation	Description
u	uninvestigated
s	simple
c	compound
a	related to assignment
i	isomorphic
e	elected from isomorphic
$\min_x$	minimum value of a given property in set x
$\max_x$	maximum value of a given property in set x

For all languages function *gc* is case-insensitive. For English and Hungarian languages function *gc* must remove all spaces and dashes from tags, because according to grammatical rules (for example number of vocals, word-class, etc.)

compound words can be written in different ways (into one, or with dashes, or with spaces). If short-length tags (which contain only some letters) exist, then dashes and spaces must be retained. Related to grammatical rules function *gc* can be determined for other languages, as well.

Table 2
Notations of algorithms of the tag recommendation system

Notation	Description
τ	set of tags
π	set of papers
α	set of assignments between tags and papers
σ	set of strings extracted from tags related to grammatical rules
t	a tag
p	a paper
l	length of a tag t
c	reference count of a tag t
s	reference count sum of a tag t
d	date and time when a tag t is created
$t_1 \mapsto t_2$	tags t_1 and t_2 are isomorphic
$t \rightarrow p$	tag t is assigned to paper p
$t\%$	starts with tag t: "t string" (string with at least 1 char)
$\%t$	ends with tag t: "string t" (string with at least 1 char)
$\%t\%$	contains tag t: "t" or "string1 t string2" (strings with at least 1 char)
$t_1 + t_2$	concatenates tag t_1 and tag t_2: "t_1 t_2"
$gc\colon \tau \rightarrow \sigma$	function related to grammatical rules in order to compare
$gr\colon \tau \rightarrow \sigma$	function related to rules to retrieve correct version

For all languages function *gr* is case-insensitive, as well. It determines related to some simple grammatical rules (for example in case of Hungarian language the number of vocals is bigger than 7) how a given tag may be written correctly.

It is worth to realize in these functions only simple grammatical rules, because the main aim of these algorithms is not spell correction based on edit distance, only vocabulary refinement by contract tags.

4.1.1 Algorithm to Correct Spelling and Clerical Errors

In tag clouds, there exist many misleading tags due to spelling and clerical errors. Spelling errors occur due to grammatical mistakes, while typing mistakes are considered as clerical errors.

In case of clerical errors it is worth to permit only one character difference, because in shorter tags two or more character difference may yield two correct but different tags. Moreover, it can be automated with minimum failure only for longer tags and for tags which do not contain numbers in specific technical environment. See for example tags "motor" and "rotor", or "IPv4" and "IPv6".

Therefore, it is more accurate to use the proposed algorithm only for collecting the potential clerical errors, and correct them under human control.

Therefore, isomorphic function can be defined related to aforementioned comments similar to function *gc*. In addition, clerical errors can be forbidden or can be permitted only for one character. The following algorithm has been provided to correct spelling and clerical errors.

Definition 1. The algorithm for *Correcting Spelling and Clerical Errors in Tag Cloud* is defined by Algorithm 1, and the associated notations are in Tables 1 and 2.

Algorithm 1. Pseudo code of the Correcting Spelling and Clerical Errors in Tag Cloud algorithm

```
$\tau_u = \tau$
while $\tau_u \neq \varnothing$ do
    for $t \in \tau_u$, where $c = c_{\max_u}$
    $\tau_i = \{\ \forall t_i \in \tau$, where $t_i \mapsto t\ \}$
    if $|\tau_i| > 1$ then
        if $|\{\ \forall t_i \in \tau_i$, where $c_i = c_{\max_i}\ \}| = 1$ then
            $t_e = t_i$, where $t_i \in \tau_i$ and $c_i = c_{\max_i}$
        else
            if $\exists t_i \in \tau_i$, where $t_i = gr(t)$ then
                $t_e = t_i$, where $t_i \in \tau_i$ and $t_i = gr(t)$
            else
                $t_e = t_i$, where $t_i \in \tau_i$ and $d_i = d_{\min_i}$
        $c_e = \sum_{\forall t_i \in \tau_i} c_i$
        $\tau_i = \tau_i \setminus \{\ t_e\ \}$
        for all $t_i \in \tau_i$ do
            $\pi_a = \{\ \forall p \in \pi$, where $\exists t_i \to p\ \}$
            for all $p \in \pi_a$ do
                $\alpha = \alpha \setminus \{\ \mathrm{t_i} \to \mathrm{p}\ \}$, $\alpha = \alpha \cup \{\ \mathrm{t_e} \to \mathrm{p}\ \}$
        $\tau = \tau \setminus \tau_\mathrm{i}$, $\tau_u = \tau_u \setminus \{\ \mathrm{t_e}\ \}$
    $\tau_u = \tau_u \setminus \tau_\mathrm{i}$
```

For a given tag, isomorphic tags are looked for. It is worth to start from the maximum reference counts of tags, because usually there are only some spelling errors compared to the number of the correct tags. So the number of iterations in the loop may decrease, namely, the computational time may become less. If there is at least one more isomorphic tag besides the given tag, firstly, try deciding which one to elect based on the reference counts, secondly, based on grammatical rules, finally, based on the identifiers. (Using automatically generated identifiers, where less identifier means earlier created tag.) The algorithm retains not in all cases the correct version of the tags, but the tags which are the same with different writing are contracted. In Step 13 of the algorithm, the reference count of this elected tag is updated. In Steps 15 to 18 assignments of the other isomorphic tags are modified for this elected tag. In Step 19 the other isomorphic tags are deleted.

On existing tag clouds this first algorithm has to be executed once in a maintenance phase. After that or for newly created tag clouds, it can be applied in two different ways: the whole algorithm is executed periodically only in maintenance phases, or only the part of Steps 4 to 19 is run when a new tag is created, an existing tag is modified, deleted. Since sets τ_i created in Step 4 of the algorithm are disjunct, it is enough to investigate the set of assignments only once after creating all sets τ_i.

4.1.2 Algorithm to Contract Tags

Some simple tags can be contracted to compound tags. For example, see the following three tags: social, network, social network. If "social network" tag exists, and tags "social" and "network" are assigned to the same papers, then for that papers these tags can be contracted to tag "social network". The contraction of tags can be an important step in improving tag clouds, because reference counts of tags can considerably be changed (see reference counts in detail in Definition 3). A novel algorithm has been provided to contract tags as follows.

Definition 2. The algorithm for *Contracting Tags in Tag Cloud* is defined by Algorithm 2, and the associated notations are in Tables 1 and 2.

Algorithm 2. Pseudo code of Contracting Tags in Tag Cloud algorithm

```
1:  $\tau_u = \tau$
2:  while $\tau_u \neq \emptyset$ do
3:      for $t_{s_1} \in \tau_u$, where $l_{s_1} = l_{\min_u}$
4:      $\tau_c = \{ \forall t_c \in \tau$, where $gc(t_c) = gc(t_{s_1}\%)$ or $gc(t_c) = gc(\%t_{s_1}) \}$
5:      for all $t_c \in \tau_c$ do
6:          if $\exists t_{s_2} \in \tau$, where $gc(t_c) = gc(t_{s_1} + t_{s_2})$ or $gc(t_c) = gc(t_{s_2} + t_{s_1})$ then
7:              $\pi_a = \{ \forall p \in \pi$, where $\exists\, t_{s_1} \to p$ and $\exists\, t_{s_2} \to p \}$
8:              for all $p \in \pi_a$ do
9:                  $c_{s_1} = c_{s_1} - 1,\ c_{s_2} = c_{s_2} - 1$
10:                 if $c_{s_1} = 0$ then $\tau = \tau \setminus \{ \mathrm{t}_{\mathrm{s}_1} \}$
11:                 if $c_{s_2} = 0$ then $\tau = \tau \setminus \{ \mathrm{t}_{\mathrm{s}_2} \}$
12:                 $c_c = c_c + 1$
13:                 $\alpha = \alpha \setminus \{ \mathrm{t}_{\mathrm{s}_1} \to \mathrm{p},\ \mathrm{t}_{\mathrm{s}_2} \to \mathrm{p} \},\ \alpha = \alpha \cup \{ \mathrm{t}_\mathrm{c} \to \mathrm{p} \}$
14:             $\tau_u = \tau_u \setminus \{ \mathrm{t}_{\mathrm{s}_1}, t_{s_2} \}$
15:         else
16:             $\tau_u = \tau_u \setminus \{ \mathrm{t}_{\mathrm{s}_1} \}$
```

For a given tag, such compound tags are looked for, which start or end with the tag, and the other parts of the compound tags are existing tags in the cloud. It is worth to start from minimum length of tags, because the chance to contract shorter tags to a longer tag are higher, so the number of iterations in the loop may be less. In Steps 8 to 13 of the algorithm, the two simple tags are contracted to the compound tag, accordingly reference counts and assignments are updated. In Steps 10 to 11 tags with zero reference count are deleted.

This second algorithm has to be executed similarly like the first one. Recall that on existing tag clouds it has to be executed once in a maintenance phase. After that or for newly created tag clouds, it can be applied in two different ways: the whole algorithm is executed periodically only in maintenance phases, or only the part of Steps 4 to 13 is run when a new tag is created, an existing tag is modified, deleted. Since the algorithm investigates tags in ascending order by length and the set of tags may be decreased during the execution, the algorithm cannot be parallelized.

4.1.3 Experimental Results

The proposed algorithms do not influence considerably the number of tags in classes. Thus, for these intermediate steps only the change in the total number of tags is presented in Table 3.

Table 3

Change in total number

Name of used algorithm	Name of total number	Hungarian	English
Original	Number of tags	5354	4199
Correct spelling and clerical errors	Number of tags Number of decrease	5304 50	4169 30
Contract tags	Number of tags Number of decrease Number of contractions	5269 35 142	4152 17 66
Improve reference counts	Number of tags Number of improvements	5269 923	4152 628

The number of decrease and contractions is not too numerous in numbers, but using Algorithm 1 such annoying spelling errors can be corrected as “Java EE” and “JavaEE", or “model-driven development” and “model driven development”, in addition, such annoying clerical errors as “optimalization” and “optimalisation”, or “activity of daily living” and “activity od daily living”.

Moreover using Algorithm 2 such coherent tags can be contracted as “ASP.NET” and “MVC” to “ASP.NET MVC”, or “IBM” and “WebSphere” to “IBM WebSphere”, or “social” and “network” to “social network”.

4.2 Reference Count Enhancement

In the second step the reference counts are enhanced in order to reflect much more effectively the reality.

4.2.1 Algorithm to Enhance Reference Counts

In tag clouds reference counts of tags are one of the most important consideration, because there is a huge number of tags in the clouds, and popular tags can be emphasized according to their reference counts.

Definition 3. The *reference count* of tag t is $|\{\ \forall p \in \pi$, where $\exists\ t \rightarrow p\ \}|$.

The reference count of a given tag can be improved to the sum of reference counts of all tags which contain the given tag. For example, see the following tags: mobile, mobile application development, mobile communication, mobile network, mobile platform, mobile robot, mobile technology, etc. All mentioned tags contain the word "mobile". Thus, the reference count of tag "mobile" can be improved to the sum of reference counts of all mentioned tags. The following algorithm has been proposed to improve reference counts.

Definition 4. The algorithm for *Improving Reference Counts in Tag Cloud* is defined by Algorithm 3, and the associated notations are in Tables 1 and 2.

Algorithm 3. Pseudo code of Improving Reference Counts in Tag Cloud algorithm

```
for all $t \in \tau$ do
        $\tau_c = \{\ \forall t_c \in \tau$, where $gc(t_c) = gc(\%t\%)\ \}$
        $s = \sum_{\forall t_c \in \tau_c} c_c$
```

It is worth storing the improved reference counts of tags (in database or cache). On existing tag clouds this third algorithm has to be executed once in a maintenance phase. After that or for newly created tag clouds, it has to be run when a new tag is created, an existing tag is modified, deleted only on tags related to the created, modified, deleted tag.

4.2.2 Experimental Results

The algorithm to improve reference count does not modify the number of tags, however it influences considerably the reference counts. In Table 3 the total number of improvements related to reference counts can be seen. The change in the reference counts can be noticed for the whole interval in Fig. 1, and from reference count 10 in Fig. 2. It can be seen on the figures that with the improvement the count of tags referenced only 1 is substantially decreased, in addition, for larger counts there are more bins (with non zero count), moreover, the counts are considerably increased.

4.3 Enriched Visualization

In the third step the font distribution algorithm is improved to efficiently divide tags to classes and easily identify popular tags.

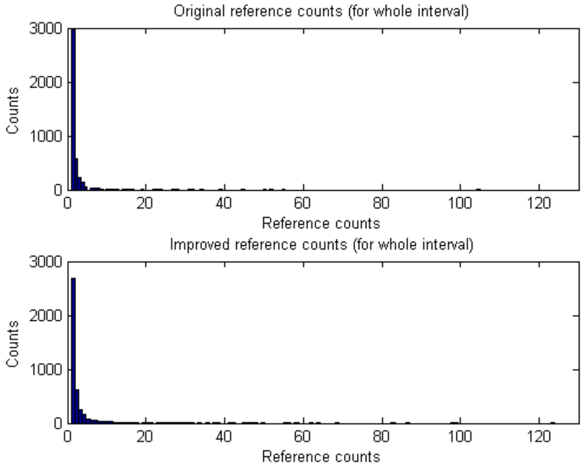


Figure 1

Histogram of original and improved reference counts for whole interval

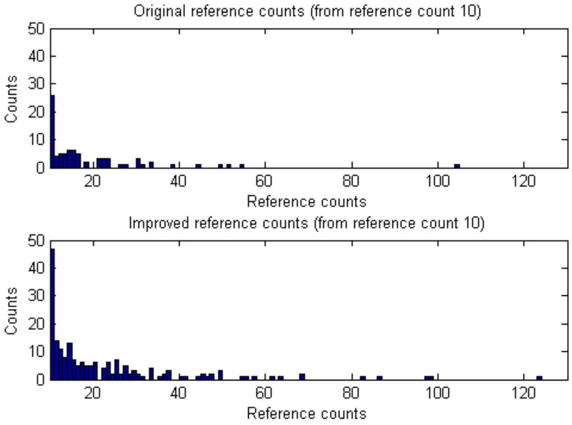


Figure 2

Histogram of original and improved reference counts starting from reference count 10

4.3.1 Distribution of Reference Counts

It seems from the histograms that the improved reference counts obey a power law. The investigation has been performed with MATLAB functions provided by [21]. Firstly, a power law distribution has been fitted to the data set of reference counts, and its parameters have been estimated. The maximum likelihood estimate of the scaling exponent is $\gamma = 2.12$, and the estimate of the lower bound of the power law behaviour is $x_{\min} = 1$. The uncertainties in the estimated parameters are 0.0185 for γ and 0 for $x_{\min}$.

Then, the test whether the power law is a plausible fit to the empirical data set of reference counts has been performed graphically and numerically, as well. The result of the graphical method shows that the empirical data follows a straight line but departs from it at the end. This means that the empirical data has a longer tail than the estimated theoretical power law distribution. The test can be performed numerically with the help of a hypothesis test using Kolmogorov-Smirnov statistic. Since the calculated p-value 0.136 is greater than 0.1 (in addition, $x_{\min} = 1$ and the number of reference counts is more than 4000), the power law is a plausible hypothesis for the data.

4.3.2 Font Distribution Algorithm

In tag clouds the tags are classified according to their reference counts. The number of classes is an arbitrary parameter. In our clouds, tags are classified to four classes, and the improved reference count sums *s* are used instead of reference counts *r*. The result of different distribution algorithms can be depicted in Table 4.

Table 4

Number of tags belonging to classes

Font distribution algorithm	Class 1	Class 2	Class 3	Class 4
Linear for *r*	4151	8	0	1
Linear for *s*	4124	27	6	3
Logarithmic for *s*	3542	471	121	26
Power law and percentage based for *s*	3935 (95%)	142 (3%)	63 (1.5%)	20 (0.5%)

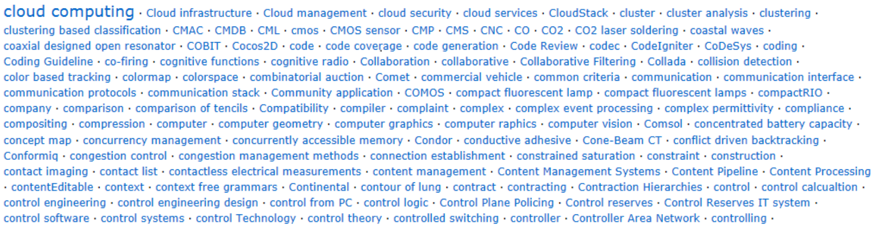
The linear distribution algorithm simply divides linearly the whole range (from minimum count to maximum) count by the number of classes. The logarithmic algorithm divides logarithmically equal intervals.

Since the improved reference counts obey a power law, a power law and percentage based approach can led to correct visual impression. The proposed distribution algorithm divides the area of the given power law function (Eq. 1) according to arbitrary percentages taking into consideration the bounds of same reference counts.

$$\frac{\max(s)^{-\gamma+1} - \min(s)^{-\gamma+1}}{-\gamma+1} \tag{1}$$

4.3.3 Experimental Results

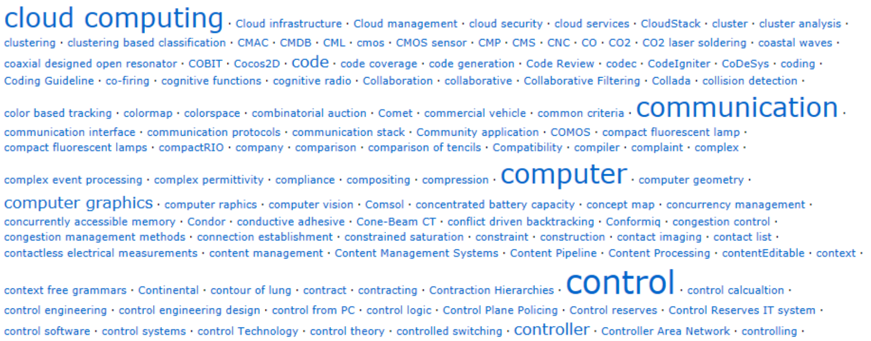
In Fig. 3 a part of the original tag cloud is illustrated. In Fig. 4 the same part of the resulted tag cloud is depicted.



cloud computing · Cloud infrastructure · Cloud management · cloud security · cloud services · CloudStack · cluster · cluster analysis · clustering ·
clustering based classification · CMAC · CMDB · CML · cmos · CMOS sensor · CMP · CMS · CNC · CO · CO2 · CO2 laser soldering · coastal waves ·
coaxial designed open resonator · COBIT · Cocos2D · code · code coverage · code generation · Code Review · codec · CodeIgniter · CoDeSys · coding ·
Coding Guideline · co-firing · cognitive functions · cognitive radio · Collaboration · collaborative · Collaborative Filtering · Collada · collision detection ·
color based tracking · colormap · colorspace · combinatorial auction · Comet · commercial vehicle · common criteria · communication · communication interface ·
communication protocols · communication stack · Community application · COMOS · compact fluorescent lamp · compact fluorescent lamps · compactRIO ·
company · comparison · comparison of tencils · Compatibility · compiler · complaint · complex · complex event processing · complex permittivity · compliance ·
compositing · compression · computer · computer geometry · computer graphics · computer raphics · computer vision · Comsol · concentrated battery capacity ·
concept map · concurrency management · concurrently accessible memory · Condor · conductive adhesive · Cone-Beam CT · conflict driven backtracking ·
Conformiq · congestion control · congestion management methods · connection establishment · constrained saturation · constraint · construction ·
contact imaging · contact list · contactless electrical measurements · content management · Content Management Systems · Content Pipeline · Content Processing
· contentEditable · context · context free grammars · Continental · contour of lung · contract · contracting · Contraction Hierarchies · control · control calcualtion ·
control engineering · control engineering design · control from PC · control logic · Control Plane Policing · Control reserves · Control Reserves IT system ·
control software · control systems · control Technology · control theory · controlled switching · controller · Controller Area Network · controlling ·

Figure 3

A part of the original tag cloud



cloud computing · Cloud infrastructure · Cloud management · cloud security · cloud services · CloudStack · cluster · cluster analysis ·
clustering · clustering based classification · CMAC · CMDB · CML · cmos · CMOS sensor · CMP · CMS · CNC · CO · CO2 · CO2 laser soldering · coastal waves ·
coaxial designed open resonator · COBIT · Cocos2D · code · code coverage · code generation · Code Review · codec · CodeIgniter · CoDeSys · coding ·
Coding Guideline · co-firing · cognitive functions · cognitive radio · Collaboration · collaborative · Collaborative Filtering · Collada · collision detection ·
color based tracking · colormap · colorspace · combinatorial auction · Comet · commercial vehicle · common criteria · communication ·
communication interface · communication protocols · communication stack · Community application · COMOS · compact fluorescent lamp ·
compact fluorescent lamps · compactRIO · company · comparison · comparison of tencils · Compatibility · compiler · complaint · complex ·
complex event processing · complex permittivity · compliance · compositing · compression · computer · computer geometry ·
computer graphics · computer raphics · computer vision · Comsol · concentrated battery capacity · concept map · concurrency management ·
concurrently accessible memory · Condor · conductive adhesive · Cone-Beam CT · conflict driven backtracking · Conformiq · congestion control ·
congestion management methods · connection establishment · constrained saturation · constraint · construction · contact imaging · contact list ·
contactless electrical measurements · content management · Content Management Systems · Content Pipeline · Content Processing · contentEditable · context ·
context free grammars · Continental · contour of lung · contract · contracting · Contraction Hierarchies · control · control calcualtion ·
control engineering · control engineering design · control from PC · control logic · Control Plane Policing · Control reserves · Control Reserves IT system ·
control software · control systems · control Technology · control theory · controlled switching · controller · Controller Area Network · controlling ·

Figure 4

A part of the resulted tag cloud

It can be seen that in the original cloud almost all tags are in unite, however in the resulted cloud many popular tags are emphasized, thus, the resulted tag cloud is much more perspicuous, the popular tags can be indentified easily.

5 Topic Recommendation System

The aim of the proposed topic recommendation system is to detect the most popular topics from the huge tag cloud.

The provided system has three main steps. In the first step a special graph is constructed from tags. In the second step the reference count of each node is evaluated in order to identify topics. In the third step the improved font distribution algorithm of the proposed tag recommendation system is applied and a visualization is introduced. Table 5 summarizes the notations of algorithms.

Table 5
Notations of algorithms of the topic recommendation system

Notation	Description
τ	set of tags
τ_u	set of uninvestigated tags
t	a tag
$t.dn$	display name property of tag *t*
$t.rc$	reference count property of tag *t*
$t.wc$	word count property of tag *t*
$dn.w_i$	i^{th} word of a display name
$dn.w_{i \to j}$	concatenated word starting from i^{th} word and ending at j^{th} word of a display name *dn*
ν	set of nodes
ε	set of edges
v	a node
v_s	source node of a directed edge
v_t	target node of a directed edge
$e_{i,j}$	an edge from node v_i to node v_j
$v.id$	identifier property of node *v*
$v.dn$	display name property of node *v*
$v.wt$	weight property of node *v*
$v.rc$	reference count property of node *v*

5.1 Graph Construction

In the first step a special graph is constructed from tags.

5.1.1 Algorithm to Construct Graph from Tags

A directed, weighted graph $G = (\nu, \varepsilon)$ is defined as a set of nodes ν and edges ε. The ij^{th} entry of the adjacency matrix *A* is 1 if there is a directed edge from node *i* to node *j*, and 0 if such edge does not exist. Only the nodes have nonnegative weights, the edges are not weighted. This graph is constructed using Algorithm 4.

Definition 5. The algorithm for *Constructing Graph from Tags* is defined by Algorithm 4, and the associated notations are in Table 5.

Algorithm 4. Pseudo code of Constructing Graph from Tags algorithm

```
$\tau_u = \tau$
$\upsilon = \emptyset,\ \varepsilon = \emptyset$
$k = 1$
while $\tau_u \neq \emptyset$ do
    for $t \in \tau_u$, where $t.wc = \min(t_u.wc),\ \forall t_u \in \tau_u$
    $k = AddNode(k, t.dn, t.wc, t.rc) + 1$

function $AddNode(k, dn, wc, rc) : i$
    if $wc = 1$ then
        $v.id = k, v.dn = dn, v.wt = rc, \upsilon = \upsilon \cup \{ v \}$
        return $v.id$
    else
        $i_1 = 0$
        for $i = 1 \to wc - 1$ do
            if $i_1 = 0$ then
                if $\exists v_1 \in \upsilon$, where $v_1.dn = dn.w_{1 \to wc-i}$ then
                    $i_1 = v_1.id$, where $v_1.dn = dn.w_{1 \to wc-i}$
                    if $\exists v_2 \in \upsilon$, where $v_2.dn = dn.w_{wc-i+1 \to wc}$ then
                        $i_2 = v_2.id$, where $v_2.dn = dn.w_{wc-i+1 \to wc}$
                    else
                        $i_2 = AddNode(k, dn.w_{wc-i+1 \to wc}, i, 0)$
                        $k = i_2 + 1$
                else if $\exists v_1 \in \upsilon$, where $v_1.dn = dn.w_{i+1 \to wc}$ then
                    $i_1 = v_1.id$, where $v_1.dn = dn.w_{i+1 \to wc}$
                    if $\exists v_2 \in \upsilon$, where $v_2.dn = dn.w_{1 \to i}$ then
                        $i_2 = v_2.id$, where $v_2.dn = dn.w_{1 \to i}$
                    else
                        $i_2 = AddNode(k, dn.w_{1 \to i}, i, 0)$
                        $k = i_2 + 1$
        if $i_1 = 0$ then
            $v_1.id = k, v_1.dn = dn.w_1, v_1.wt = 0, \upsilon = \upsilon \cup \{ v_1 \}$
            $i_1 = v_1.id$
            $i_2 = AddNode(k + 1, dn.w_{2 \to wc}, wc - 1, 0)$
            $k = i_2 + 1$
        $v.id = k, v.dn = dn, v.wt = rc, \upsilon = \upsilon \cup \{ v \}$
        $\varepsilon = \varepsilon \cup \{ e_{i_1, v.id}, e_{i_2, v.id} \}$
        return $v.id$
```

This is a recursive algorithm. The tags are investigated ordered ascending by their word count. If the word count is one, then create a new node for it, and return with their identifier in Steps 9-11.

If the display name of the tag contains more words, then search for a matching node investigating the display name from its begin (Steps 16-17) and its end (Steps 23-24) starting from length of word count minus one until one. If a matching node exists, then investigate whether the other part of the display name exists as a node (Steps 18-19, 25-26).

If there is no matching for the other part, then create appropriate nodes for it by recursive calls (Steps 21-22, 28-29). If there is no matching node investigating the display name from its begin and its end starting from length of word count minus one until one (Step 30), then create a new node for the first word (Steps 31-32), and create appropriate nodes for the other part of the display name by recursive calls (Steps 33-34).

After the shorter parts of the display name exist, then create a new node for the whole display name (Step 35). Moreover, create two directed edges from two nodes with shorter parts to the node of the whole display name (Step 36).

5.1.2 Experimental Results

An example part of the constructed tag graph can be seen in Fig. 5. The tags are the following: data, text mining, data mining, data mining competition.

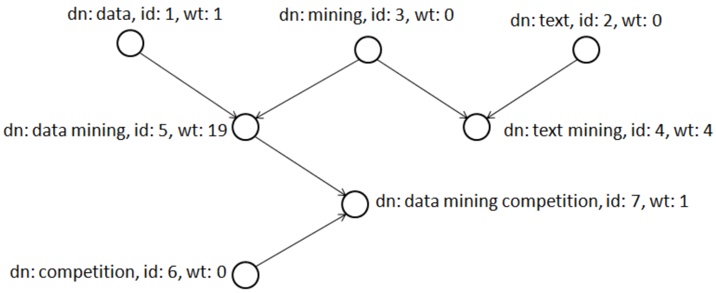

dn: data, id: 1, wt: 1
dn: mining, id: 3, wt: 0
dn: text, id: 2, wt: 0
dn: data mining, id: 5, wt: 19
dn: text mining, id: 4, wt: 4
dn: data mining competition, id: 7, wt: 1
dn: competition, id: 6, wt: 0

Figure 5
An example part of the tag graph

The reference counts of such nodes which do not correspond to real tags are zero (Steps 21, 28, 33 of Algorithm 4). The in-degree of nodes can be zero or two: zero if the display name of a given node contains only one word, in addition, two if it is a compound word.

The histogram of out-degree of nodes is depicted in Fig. 6. There are numerous nodes whose out-degree is zero, namely, they are not building items of nodes with longer display name. However, there are some nodes which are frequently used building items. The weight and out-degree of nodes influences the reference counts of nodes, namely, the recommended topics.

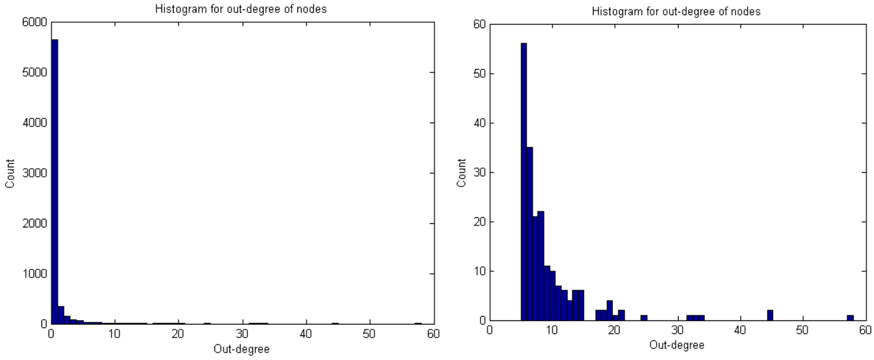


Figure 6

Histogram for out-degree of nodes (left: all, right: only from out-degree 5)

5.2 Topic Identification

In the second step the reference count of each node is evaluated in order to identify topics.

5.2.1 Algorithm to Recommend Topics from Graph

In the second step the reference counts of nodes are calculated by Algorithm 5. The reference count of a node is calculated as the sum of weights of such nodes which can be reached from the given node via directed edges. The proposed special tag graph with the calculated reference counts can be used for topic recommendation. The recommended topics are the nodes with the most reference counts. The limit in reference counts or the maximum number of topics can be chosen arbitrary.

Definition 6. The algorithm for *Calculating Reference Counts of Nodes* is defined by Algorithm 5, and the associated notations are in Table 5.

Algorithm 5. Pseudo code of *Calculating Reference Counts of Nodes* algorithm

```
for all $\upsilon \in \upsilon$ do
    $v.rc = v.wt + Count(v,0)$
function $Count(v_s, rc) : rc$
    $\upsilon_t = \{\ v_t$, where $\exists e_{s,t}\ \}$
    if $|\upsilon_t| \geq 1$ then
        $rc = rc + \sum_{\forall v_t \in \upsilon_t} v_t.wt$
        for all $v_t \in \upsilon_t$ do
            $rc = Count(v_t, rc)$
    return $rc$
```

5.2.2 Experimental Results

The construction of the tag graph is described in Table 6.

Table 6

Construction of tag graph

	Count
Tags	4152
Nodes	6408
Edges	5404
Nodes whose in-degree = 0	3706
Nodes whose in-degree = 2	2702
$rc \geq 10$	274
$wt \neq rc$ ($rc \geq 10$)	236
$wt = 0$ ($rc \geq 10$)	70

The nodes whose reference count is greater or equal to 10 are identified as topics. The reference counts and the weights of topics are different numbers in more than 85 percentages, thus, the proposed reference counts of nodes are an important improvement of the reference count of tags. More than 25 percentages of nodes are such topics, which are not existing tags, hence, the provided tag graph is a significant enhancement of the original tag cloud.

5.3 Visualization of Recommended Topics

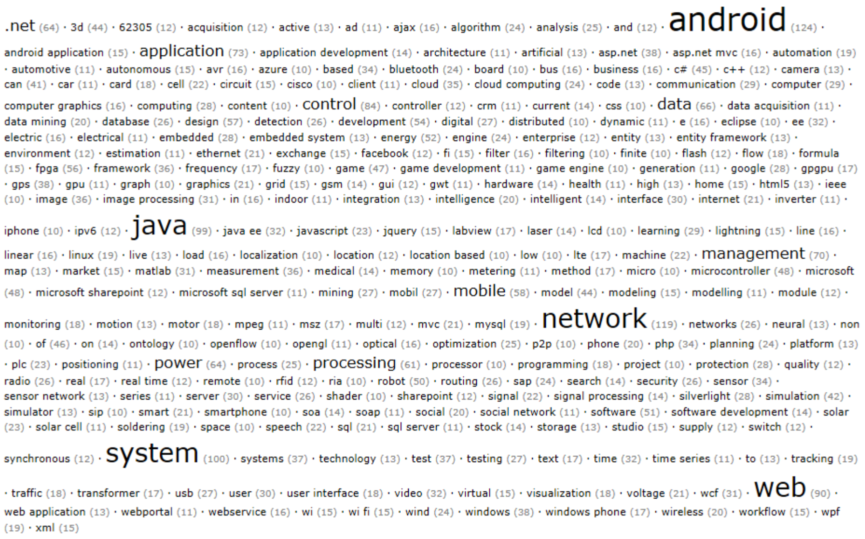
In the third step the improved font distribution algorithm of the proposed tag recommendation system (see Section 4.3) is applied and a visualization is introduced.

The resulted distribution of nodes among classes is summarized in Table 7. The resulted topic cloud are depicted in Figs. 7 and 8.

Table 7

Distribution of nodes among classes

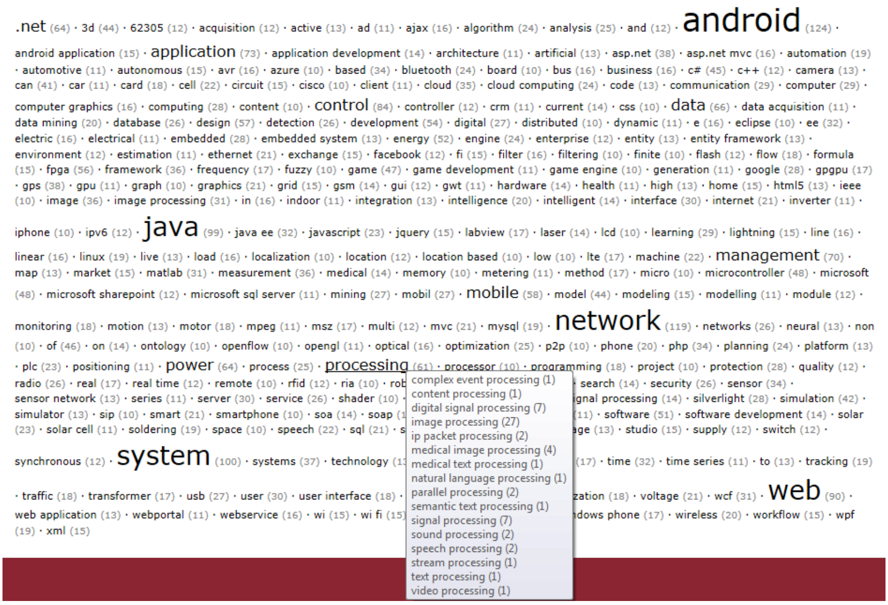
Class	Percentage	Count of nodes
Class 1	95	255
Class 2	3	8
Class 3	1.5	4
Class 4	0.5	1



.net (64) · 3d (44) · 62305 (12) · acquisition (12) · active (13) · ad (11) · ajax (16) · algorithm (24) · analysis (25) · and (12) · android (124) ·
android application (15) · application (73) · application development (14) · architecture (11) · artificial (13) · asp.net (38) · asp.net mvc (16) · automation (19)
· automotive (11) · autonomous (15) · avr (16) · azure (10) · based (34) · bluetooth (24) · board (10) · bus (16) · business (16) · c# (45) · c++ (12) · camera (13) ·
can (41) · car (11) · card (18) · cell (22) · circuit (15) · cisco (10) · client (11) · cloud (35) · cloud computing (24) · code (13) · communication (29) · computer (29) ·
computer graphics (16) · computing (28) · content (10) · control (84) · controller (12) · crm (11) · current (14) · css (10) · data (66) · data acquisition (11) ·
data mining (20) · database (26) · design (57) · detection (26) · development (54) · digital (27) · distributed (10) · dynamic (11) · e (16) · eclipse (10) · ee (32) ·
electric (16) · electrical (11) · embedded (28) · embedded system (13) · energy (52) · engine (24) · enterprise (12) · entity (13) · entity framework (13) ·
environment (12) · estimation (11) · ethernet (21) · exchange (15) · facebook (12) · fi (15) · filter (16) · filtering (10) · finite (10) · flash (12) · flow (18) · formula
(15) · fpga (56) · framework (36) · frequency (17) · fuzzy (10) · game (47) · game development (11) · game engine (10) · generation (11) · google (28) · gpgpu (17)
· gps (38) · gpu (11) · graph (10) · graphics (21) · grid (15) · gsm (14) · gui (12) · gwt (11) · hardware (14) · health (11) · high (13) · home (15) · html5 (13) · ieee
(10) · image (36) · image processing (31) · in (16) · indoor (11) · integration (13) · intelligence (20) · intelligent (14) · interface (30) · internet (21) · inverter (11) ·
iphone (10) · ipv6 (12) · java (99) · java ee (32) · javascript (23) · jquery (15) · labview (17) · laser (14) · lcd (10) · learning (29) · lightning (15) · line (16) ·
linear (16) · linux (19) · live (13) · load (16) · localization (10) · location (12) · location based (10) · low (10) · lte (17) · machine (22) · management (70) ·
map (13) · market (15) · matlab (31) · measurement (36) · medical (14) · memory (10) · metering (11) · method (17) · micro (10) · microcontroller (48) · microsoft
(48) · microsoft sharepoint (12) · microsoft sql server (11) · mining (27) · mobil (27) · mobile (58) · model (44) · modeling (15) · modelling (11) · module (12) ·
monitoring (18) · motion (13) · motor (18) · mpeg (11) · msz (17) · multi (12) · mvc (21) · mysql (19) · network (119) · networks (26) · neural (13) · non
(10) · of (46) · on (14) · ontology (10) · openflow (10) · opengl (11) · optical (16) · optimization (25) · p2p (10) · phone (20) · php (34) · planning (24) · platform (13)
· plc (23) · positioning (11) · power (64) · process (25) · processing (61) · processor (10) · programming (18) · project (10) · protection (28) · quality (12) ·
radio (26) · real (17) · real time (12) · remote (10) · rfid (12) · ria (10) · robot (50) · routing (26) · sap (24) · search (14) · security (26) · sensor (34) ·
sensor network (13) · series (11) · server (30) · service (26) · shader (10) · sharepoint (12) · signal (22) · signal processing (14) · silverlight (28) · simulation (42) ·
simulator (13) · sip (10) · smart (21) · smartphone (10) · soa (14) · soap (11) · social (20) · social network (11) · software (51) · software development (14) · solar
(23) · solar cell (11) · soldering (19) · space (10) · speech (22) · sql (21) · sql server (11) · stock (14) · storage (13) · studio (15) · supply (12) · switch (12) ·
synchronous (12) · system (100) · systems (37) · technology (13) · test (37) · testing (27) · text (17) · time (32) · time series (11) · to (13) · tracking (19)
· traffic (18) · transformer (17) · usb (27) · user (30) · user interface (18) · video (32) · virtual (15) · visualization (18) · voltage (21) · wcf (31) · web (90) ·
web application (13) · webportal (11) · webservice (16) · wi (15) · wi fi (15) · wind (24) · windows (38) · windows phone (17) · wireless (20) · workflow (15) · wpf
(19) · xml (15)

Figure 7

Resulted topic cloud



.net (64) · 3d (44) · 62305 (12) · acquisition (12) · active (13) · ad (11) · ajax (16) · algorithm (24) · analysis (25) · and (12) · android (124) ·
android application (15) · application (73) · application development (14) · architecture (11) · artificial (13) · asp.net (38) · asp.net mvc (16) · automation (19)
· automotive (11) · autonomous (15) · avr (16) · azure (10) · based (34) · bluetooth (24) · board (10) · bus (16) · business (16) · c# (45) · c++ (12) · camera (13) ·
can (41) · car (11) · card (18) · cell (22) · circuit (15) · cisco (10) · client (11) · cloud (35) · cloud computing (24) · code (13) · communication (29) · computer (29) ·
computer graphics (16) · computing (28) · content (10) · control (84) · controller (12) · crm (11) · current (14) · css (10) · data (66) · data acquisition (11) ·
data mining (20) · database (26) · design (57) · detection (26) · development (54) · digital (27) · distributed (10) · dynamic (11) · e (16) · eclipse (10) · ee (32) ·
electric (16) · electrical (11) · embedded (28) · embedded system (13) · energy (52) · engine (24) · enterprise (12) · entity (13) · entity framework (13) ·
environment (12) · estimation (11) · ethernet (21) · exchange (15) · facebook (12) · fi (15) · filter (16) · filtering (10) · finite (10) · flash (12) · flow (18) · formula
(15) · fpga (56) · framework (36) · frequency (17) · fuzzy (10) · game (47) · game development (11) · game engine (10) · generation (11) · google (28) · gpgpu (17)
· gps (38) · gpu (11) · graph (10) · graphics (21) · grid (15) · gsm (14) · gui (12) · gwt (11) · hardware (14) · health (11) · high (13) · home (15) · html5 (13) · ieee
(10) · image (36) · image processing (31) · in (16) · indoor (11) · integration (13) · intelligence (20) · intelligent (14) · interface (30) · internet (21) · inverter (11) ·
iphone (10) · ipv6 (12) · java (99) · java ee (32) · javascript (23) · jquery (15) · labview (17) · laser (14) · lcd (10) · learning (29) · lightning (15) · line (16) ·
linear (16) · linux (19) · live (13) · load (16) · localization (10) · location (12) · location based (10) · low (10) · lte (17) · machine (22) · management (70) ·
map (13) · market (15) · matlab (31) · measurement (36) · medical (14) · memory (10) · metering (11) · method (17) · micro (10) · microcontroller (48) · microsoft
(48) · microsoft sharepoint (12) · microsoft sql server (11) · mining (27) · mobil (27) · mobile (58) · model (44) · modeling (15) · modelling (11) · module (12) ·
monitoring (18) · motion (13) · motor (18) · mpeg (11) · msz (17) · multi (12) · mvc (21) · mysql (19) · network (119) · networks (26) · neural (13) · non
(10) · of (46) · on (14) · ontology (10) · openflow (10) · opengl (11) · optical (16) · optimization (25) · p2p (10) · phone (20) · php (34) · planning (24) · platform (13)
· plc (23) · positioning (11) · power (64) · process (25) · processing (61) · processor (10) · programming (18) · project (10) · protection (28) · quality (12) ·
radio (26) · real (17) · real time (12) · remote (10) · rfid (12) · ria (10) · rob
search (14) · security (26) · sensor (34) ·
sensor network (13) · series (11) · server (30) · service (26) · shader (10) ·
ignal processing (14) · silverlight (28) · simulation (42) ·
simulator (13) · sip (10) · smart (21) · smartphone (10) · soa (14) · soap (1
(11) · software (51) · software development (14) · solar
(23) · solar cell (11) · soldering (19) · space (10) · speech (22) · sql (21) · s
age (13) · studio (15) · supply (12) · switch (12) ·
synchronous (12) · system (100) · systems (37) · technology (1
(17) · time (32) · time series (11) · to (13) · tracking (19)
· traffic (18) · transformer (17) · usb (27) · user (30) · user interface (18) ·
zation (18) · voltage (21) · wcf (31) · web (90) ·
web application (13) · webportal (11) · webservice (16) · wi (15) · wi fi (15)
dows phone (17) · wireless (20) · workflow (15) · wpf
(19) · xml (15)

complex event processing (1)
content processing (1)
digital signal processing (7)
image processing (27)
ip packet processing (2)
medical image processing (4)
medical text processing (1)
natural language processing (1)
parallel processing (2)
semantic text processing (1)
signal processing (7)
sound processing (2)
speech processing (2)
stream processing (1)
text processing (1)
video processing (1)

Figure 8

Resulted topic cloud with tooltip about tags belonging to node of "processing"

The identified topics are visualized as a tag cloud alphabetically ordered and visually weighted by letter size. The calculated reference counts are in brackets after the display name of topics. The tags with their reference counts, which fill parts of the given node, are shown on tooltips. In Fig. 8 see for example the tooltip for node 'processing', who is not in the original tag cloud, who is not an itself existing tag, but identified as an important topic.

Conclusions

The visualization of huge tag clouds is one of the most important and complicated consideration. In our thesis portal tag clouds have a very important role to facilitate browsing and searching among numerous tags and theses. In this paper complete tag and topic recommendation systems have been provided.

The tag recommendation system has three steps. In the first step the vocabulary has been refined. The aim of the algorithm correcting spelling and clerical errors is not to correct all spelling and clerical errors in each tag, but to contract such tags which are the same only with different writing (the algorithm retains not in all cases the grammatically correct version). The purpose of the algorithm contracting tags is to contract such simple tags of each thesis to a compound tag in which case this compound tag is used for another thesis. In the second step the reference counts have been enhanced in order to reflect much more effectively the reality. In the third step the font distribution algorithm has been improved to efficiently divide tags to classes and easily identify popular tags. Using the established tag recommendation system the quality of tags, the detection of the actually popular tags, and the visualization of tag clouds can be improved.

The topic recommendation system has three steps. In the first step a special graph has been constructed from tags. In the second step the reference count of each node has been evaluated in order to identify topics. In the third step the improved font distribution algorithm of the proposed tag recommendation system has been applied and a visualization has been introduced. The resulted topic cloud is a significant enhancement of the original tag cloud, since lots of such topics are identified, which are not existing tags in the original tag cloud, in addition, the popularity of topics is evaluated more properly, furthermore, popular topics can be detected easily.

The proposed systems have been implemented in C# classes using LINQ, SQL stored procedures, and MATLAB functions. They have been validated and verified on tag clouds of a real-world thesis portal.

Acknowledgement

First of all we wish to thank SUZUKI Foundation for sponsoring this research. This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013).

References

[1] T. O'Reilly, What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software, International Journal of Digital Economics, No. 65, pp. 17-37, March 2007

[2] A. M. Kaplan, M. Haenlein, Users of the World, Unite! The Challenges and Opportunities of Social Media, Business Horizons, Vol. 53(1), pp. 59-68, 2010

[3] J. Zimmerman, D. Sahlin, Social Media Marketing All-in-One for Dummies, Wiley Publishing, ISBN 978-0-470-58468-2, 2010

[4] C. C. Aggarwal, Social Network Data Analytics, 1st Edition, Springer, ISBN 978-1-4419-8461-6, 2011

[5] Microsoft Research - Social Computing Group, http://research.microsoft.com/scg

[6] IBM Research - Social Computing Group, http://www.research.ibm.com/haifa/dept/imt/ct st.shtml

[7] HP Labs - Social Computing Lab, http://www.hpl.hp.com/research/scl

[8] S. A. Golder, B. A. Huberman, The Structure of Collaborative Tagging Systems, 2005

[9] D. H. Pink, Folksonomy, New York Times, December 11, 2005

[10] J. Sinclair, M. Cardew-Hall, The Folksonomy Tag Cloud: When is It Useful?, Journal of Information Science, Vol. 34(1), pp. 15-29, 2008

[11] S. Lohmann, J. Ziegler, L. Tetzlaff, Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration, INTERACT, Part I, LNCS 5726, pp. 392-404, 2009

[12] Y. Hassan-Montero, V. Herrero-Solana, Improving Tag-Clouds as Visual Information Retrieval Interfaces, International Conference on Multidisciplinary Information Sciences and Technologies, Merida, Spain, 2006

[13] J. Salonen, Self-Organizing Map Based Tag Clouds, OPAALS Conference, 2007

[14] kentbye's blog, Tag Cloud Font Distribution Algorithm, http://www.echochamberproject.com/node/247, June 24, 2005

[15] K. Hoffman, In Search of ... The Perfect Tag Cloud, http://files.blog-city.com/files/J05/88284/b/insearchofperfecttagcloud.pdf

[16] A. Clauset, C.R. Shalizi, and M. E. J. Newman, Power-Law Distributions in Empirical Data, SIAM Review, Vol. 51(4), pp. 661-703, 2009

[17] Á. Bogárdi-Mészöly, A. Rövid, H. Ishikawa, Algorithms to Improve Tag Clouds, 5th International Conference on Emerging Trends in Engineering & Technology, Himeji, Japan, pp. 303-308, ISBN 978-0-7695-4884-5, 2012

[18] Á. Bogárdi-Mészöly, A. Rövid, H. Ishikawa, An Improved Font Distribution Algorithm for Tag Clouds, 6th International Conference on Soft Computing and Intelligent Systems, Kobe, Japan, pp. 2264-2267, ISSN 1880-3741, 2012

[19] Á. Bogárdi-Mészöly, A. Rövid, H. Ishikawa, Topic Recommendation from Tag Clouds, 2nd International Workshop on Networking, Computing, Systems, and Software, Okinawa, Japan, 2012, pp. 25-29, Bulletin of Networking, Computing, Systems, and Software, ISSN 2186–5140, Vol. 2(1), January 2013

[20] Thesis Portal, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, http://diplomaterv.vik.bme.hu

[21] A. Clauset, C.R. Shalizi, and M. E. J. Newman, Implementation for Article of Power-Law Distributions in Empirical Data, http://tuvalu.santafe.edu/~aaronc/powerlaws/

Business Process Modeling and the Robust PNS Problem

József Tick

Óbuda University, Bécsi út 96/b, H-1034 Budapest, Hungary
e-mail: tick@uni-obuda.hu

Csanád Imreh, Zoltán Kovács

University of Szeged, Árpád tér 2, H-6720 Szeged, Hungary
e-mail: cimreh@inf.u-szeged.hu; kovacsz@inf.u-szeged.hu

Abstract: In this paper we define and investigate a new direction of the P-graph-based Business Process Modeling which we call the robust PNS problem. We consider the model where for each operating unit two costs are given, it has a nominal cost and an extended cost, and we know that at most b operating units have the extended cost, the others will have the nominal cost. We present a branch and bound based exact solution algorithm for the general problem, and a faster, polynomial time dynamic programming algorithm for the case of the hierarchycal problems.

Keywords: Business Process Modeling; PNS problem; robust problems; P-graphs

1 Introduction

In the last decades one of the most significant changes was that informatics spread in all the fields of business processes, and has become an unavoidable component. By now business processes have become more and more complex. Their regular operations need the smooth cooperation of several systems. The solution of the cost effective and error free management of complex business processes as well as the efficiency of the solution is one of the most important issues for the profit-oriented sector, since the optimal efficiency of the complex business, production and office-automation processes is a high priority economic value.

In favor of the optimal operation modeling the business processes is necessary; furthermore, the model-based approach is necessary for the development and management of such systems.

There are more well-known approaches for the modeling of business processes. Workflow is the most widespread modeling technique in the field of industrial and office information systems. The workflow-based models can be used excellently for the analysis, modeling and optimizing of work flows in the business processes.

Several workflow representations have spread, out of which, however, the P-graph-based workflow [17] is the one that is a process-based method with correct mathematical background and gives definitely an optimal workflow network structure on a systematic way. The introduction of the p-graph in workflows has been done similarly to the modeling of process networks [18], that is why the analysis of PNS methodologies and the elaboration of further procedures are necessary.

In a manufacturing system, materials of different properties are changed by different transformation to yield desired products. Usually some raw materials and the desired products are given and our goal is to produce the desired products from the raw materials though the possible transformations. These sytems can be modelled in the P-graph framework where a bipartite graph is used. One of the sets of the vertices contains the possible materials. The other set contains the possible transformations represented as operating units defined by their input and output material sets. Some subgraphs of the graph containing all the possible transformations describe the feasible processes which produce the desired products from the raw materials and the goal is to find the cheapest such subgraph. In the combinatorial version studied in this paper each operating unit has a fixed cost and the cost of a subgraph is the sum of the costs of the operating units contained in it. In a more general, quantitive model we also consider the amounts of the used materials and the cost of the operating units depends on the amount of materials used by them. One can find the detailed backround of the PNS model in [6], [7], therefore we will recall only the main definitons in the next section.

The P-graph framework was designed to analyse and solve process network synthesis problems but later it was observed that it can be used to solve optimization problems in other areas as well. In [5] and [15] one can find how to use the P-graphs in supply chain management. Another big area where the P-graphs are useful is workflow management.

The [18], [19] and [20] present the P-graph-based modeling of workflows introduced to the analogy of process-network modeling as well as introduce the extension of the model by the time factor as special resource. It generates the relevant mathematical model from the PNS structure of the defined workflow and conducts analyses in order to define the objective function (capacity constraints, bottle necks) determined by the environment.

The [16] and [21] present the possible fuzzy extension of the P-graph-based workflow model, analyses the „real life like", uncertain and not-exactly definable situations present in the real application environment of the workflow, which for better modeling makes the fuzzy extension necessary. It introduces the fuzzy sets defined for documents and activities and presents the application of the method with an example. It introduces the parametric t-norm, which can be considered as the extension and generalization of the Zadeh and the Fodor t-norms indeed and with its application the tuning of the fuzzy regulatory systems might become more effective.

In these optimizaton problems we suppose that all costs are known exactly in advance. On the other hand, in real applications usually some uncertainty can change the data. In the PNS model these uncertainties can be handled in the case of workflow problems by fuzzy methods (see [16] for details) or at the supply chain problems by extending the P-graph with the ROA (reliability of availability) value at the materials (see [15] for details). In general, for most optimization model the problem of uncertainty is solved by stochastic optimization. On the other hand, in these cases we need some a priori information about the distribution of the data, which is usually not available in real applications. Another approach is a robust optimization, where the uncertainty is handled by deterministic worst case scenario. In these models we do not have the fixed values of the parameters we only know that they are in a given interval. There are many robust models of combinatorial problems one can find an overview in [3]. In this paper we consider the robust version where we have an a priori bound on the number of the parameters which might be changed, such models are studied in [4] and [14]. This means that we consider the model where for each operating unit two costs are given, it has a nominal cost and an extended cost, and we know that at most b operating units have the extended cost, the others will have the nominal cost. We will search for the optimal solution for this objective functions.

2 The Mathematical Backround

2.1 The Basic Definitions of the PNS Model

The structure of the PNS problem can be studied by the P-graphs defined in [7]. To define this graph let M be the finite nonempty set of the possible materials. The are two distinguished subsets of the materials, R denotes the set of raw materials, and P denotes the set of the desired products. The possible transformations are modeled as operating units, each of them is determined by two sets of materials, i.e., the set of input and output materials of the operating unit. For an operating unit u we denote the set of input materials by in(u), the set of output materials by out(u), we will use the same notation for the sets of operating units. We denote the set of all possible operating units by O. Then the process graph or P-graph in short is defined by this pair (M, O). The set of vertices of this directed graph is M ∪ O, and the set of edges consist of the following subsets

1) the edges which go to an operating unit from its input materials

2) the edges which go from an operating unit to its output materials.

Then some subgraphs of the P-graph represent the feasible solutions which are able to produce the required materials from the set of raw materials. In [7] it is shown that a subgraph (m,o), where m and o are the subsets of M and O, represents a feasible solution if and only if it satisfies the properties listed below.

(A1) P is a subset of m,

(A2) a material from m is a raw material if and only if no edge goes into it in the P-graph (m, o),

(A3) for every operating unit of o there exists a path in the P-graph (m,o) which goes from the unit into a desired product,

(A4) all of the materials in m are either input or output materials of some operating units from set o.

For an arbitrary material m we denote by $\Delta(m)$ the set of operating units which produce the material. We extend this notation to sets as well, $\Delta(S)$ denotes the set of operating units which produces some elements from set S.

In the combinatorial optimization version of the standard PNS problem each operating unit o has a cost c(o) and the goal is to find the feasible solution where the total cost of the operating units contained in it is minimal. There are some branch and bound algorithms for the solution of this optimization problem see [10] and [11] for details. These branch and bound algorithms are based on the notion of decision mappings which are defined in [8]. We will use these decision mappings later, thus we recall this definition here.

A decision mapping assigns to each material m a subset of $\Delta(m)$ denoted by $\delta(m)$ and this shows which operating units produce the material in the feasible solution considered. On the other hand, not an arbitrary set of decision mappings can belong to a solution. If an operating unit with output set containing a material X is selected to produce in a solution a material Y, then it must be also selected to produce X as well. These contradictions are eliminated in the consistent decision mappings where it is valid for any pair of materials X, Y that $\delta(X)\cap\Delta(Y)$ is a subset of $\delta(Y)$. The consistent decision mappings are important since each feasible solution can be described by these decision mappings of the materials.

2.2 The Robust Version

In the robust model each operating unit o_i has an extended cost $c(o_i)+ e(o_i)$. We will call $c(o_i)$ the nominal cost and $e(o_i)$ the extra cost of the operating units. For any set Q of operating units we will use c(Q) to denote the sum of the nominal costs in set Q. Furthermore, we have an a priori bound b, which means that b operating units can have the extended cost and the others have the nominal cost. We are interested in the worst case, therefore, if we consider a feasible solution of the problem in the robust version its cost will be the sum of the nominal costs of the operating units plus the sum of the b largest extra costs. In the robust model we can have completely different optimal solutions than in the case of nominal costs, as the following example shows.

Example 1. Suppose we have a problem where R_1 is the raw material, P_1 is the desired product and we have one further material denoted by $X_{1.}$ There are three operating units: U_1 produces directly P_1 from R_1 and $c(u_1)=5$ $e(u_i)=2$, U_2 produces X_1 from R_1 and $c(u_2)=2$ $e(u_2)=2$, U_3 produces P_1 from X_1 and $c(u_3)=2$ $e(u_3)=2$. If we consider the standard problem then the optimal solution contains U_2 and U_3 and the optimal cost is 4. If we consider the robust version with b=1, then the optimal solution still contains U_2 and U_3 and the optimal cost is 6. But if we consider the robust version with b=2, then the optimal solution contains U_1 and and the optimal cost is 7.

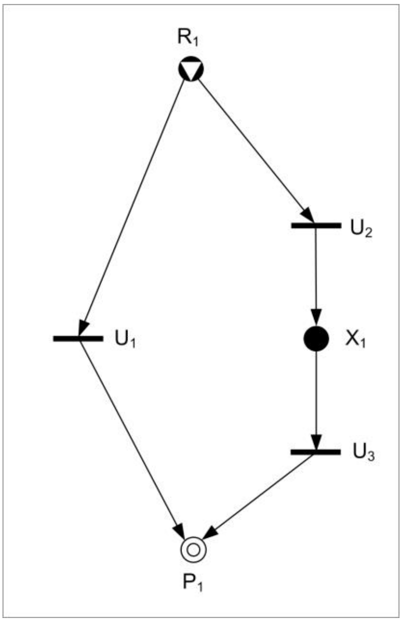


Figure 1
The P-graph of example 1

3 The Branch and Bound Algorithm

Several branch and bound-based algorithms were developed for the solution of the standard combinatorial PNS problem, one can find the details in [9] or in [11]. Here we present an extension which can be used for the the robust problem. We consider a tree where the leaves contain the feasible solutions of the problem. As it is mentioned in the previous section these solutions are identified by the decision mappings defined on the full set of materials. In the inner points of the tree we have partial solutions identified by decision mappings defined only on a subset of the materials. Since the set of the feasible solutions is the same for the standard and the robust problem our algorithm for the robust problem differs only in the bounding function, where we need an estiomation on the new cost function.

The algorithm uses a bounding function B which is defined on all partially defined decision mapping. It gives a lower bound on the costs of the feasible solutions which are the extensions of the decision mapping. The simplest function contains the total nominal cost of the selected operating units plus the sum of the b greatest extra costs among them. Some more difficult bounding functions are presented in [9] and [11], it is an interesting further question to extend them into the robust model.

Algorithm B and B

Initialization: Let the root of the tree be the empty decision mapping, calculate its bounding function and let the list L of the actual nodes contain this root with this value. Moreover, let OPT=N where N is greater than the sum of the extended cost, and OPTG be the empty graph. Continue the procedure with the iteration part.

Iteration Part (while L contains some elements):

Step 1. Choose the element of L, which has the smallest bounding function value, denote it by δ. Consider a material Y where the decision mapping δ is not defined. Extend the decision mapping of δ with all the possible consistent value of Y, denote the possible extensions by $\delta_{1,\ldots,}$ δ_t.

Step 2. If δ_i belongs to a feasible solution for some i, and the cost of the solution is smaller than OPT, then change OPTG with δ_i and OPT with the cost of δ_i.

Step 3. Delete δ from the list L and put all δ_i which satisfies $B(\delta_i) \geq$OPT into L.

Solution: The decision mapping of an optimal solution is stored in OPTG and the optimal value is stored in OPT.

We note that we might have a faster algorithm if we start with a solution given by some heuristic algorithms. There are some heuristic developed for the solution of the standard PNS problem (see [2]) and since the set of feasible solutions is the same for the standard and the robust problem they give a feasible solution for the robust problem as well. On the other hand, it would be interesting to modify these algorithms or design new heuristics for the solution of the robust PNS problem.

4 The Hierarchical Robust PNS Problems

There is a polynomial time solvable class of the PNS problem, which was investigated in [1] and [12]. In this section we extend this algorithm to the more general robust PNS problem. First we recall the definitions. A PNS problem is called *hierarchical* if there exist such partition M_0=R,...,M_l=P of M and partition O_1,..., O_l of O that for each i, i=1,...,l O_i contains operating unit having input materials from M_{i-1} and output materials from M_i. A hierarchycal PNS problem is called *k-wide hierarchical* if $M_i \vert \leq$ k is valid for i=0,...,l and $\vert O_i \vert \leq$ k is valid for i=1,....,l. Figure 2 shows a 3-wide hierarchical problem where M_0={ $R_{1,}$ R_2}, O_1={$U_{1,}$ $U_{2,}U_3$}, M_1={$X_{1,}$ $X_{2,}X_3$}, O_2={$U_{4,}$ U_5}, M_2={R_1}.

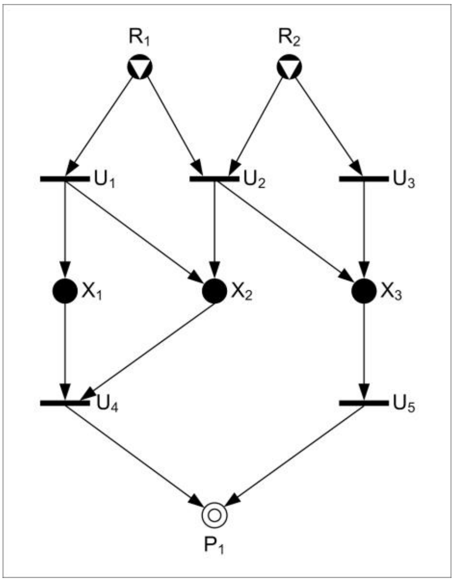


Figure 2
A 3-wide hierarchical P-graph

The assumption that a PNS problem has hierarchical structure seems to be very strong. On the other hand such problems might appear in practical applications. One example is an economic process that works by scheduled order. Then the P-graph contains no cycle and can form a hirarchical structure.

We can solve the hierarchical robust PNS problems with the following algorithm. The basic idea is to define the function F(S,j) which is the cost of the optimal solution which can produce the set S with at most i operating units having extended costs. We calculate this value by dynamic programming for each j=0,...,b and for each subsets of the sets $M_{i.}$ Then F(P,b) gives the cost of the optimal solution of the problem. We will use an other, set valued function denoted by G(S,j) which notes that in the optimal solution of F(S,j) which set of operating units is used to directly produce S. Finally V(S,j) denotes the number of the operating units which have extended cost among the operating units of G(S,j) in this optimal solution.

Algorithm HSolve

Initialization Let N be a large number which is greater than the sum of the extended costs of the machines, we will use that number to observe the cases where the problem has no feasible solution. Moreover, let F(S,j)=0 for each subset S of R= M_0 and for each j.

The i-th iteration (i=1,...,l) of computing the optimal cost Execute the following steps for each subset S of M_i and for each j=0,...,b.

Step 1. Consider the subsets of Delta(S) (note that each of these sets is included in O_i) and for each such set Q examine that the union of the output materials includes S or not. Denote the sets where S is a subset of that union by $Q_{1,...,}$ Q_t. If no such set exists then F(S,j)=N, V(S,j)=N, and G(S,j) is empty for each j. Otherwise go to step 2.

Step2. For each Q_r, r=1,...t let q_r denote the number of the operating units in the set and $E(Q_r,p)$ the sum of the p largest extra costs in set Q_r for each p=0,...,q_r. Now calculate the following value for each r=1,....,t.

$$C_r = \max\{ F(in(Q_r),j-p) + C(Q_r) + E(Q_r,p) \mid p=0,..., \min(j, q_r) \}$$

Let the set with the minimal value be denoted by Q, the minimal value is by C, the value of p where the minimum is obtained by p*. If there are more minimal values choose the set with the smallest index. Then let F(S,j)=C, G(S,j)=Q, N(S,j)=p*.

Determination of the optimal structure If F(P,b)≥N, then the problem has no feasible solution. Otherwise let A=P, v=b and O* be empty, and perform the following step while A is not a subset of the raw materials. Let O*= O* ∪ G(A,v) and let A=in(G(A,v)) and let v=v−V(A,j).

The optimal solution is the P-graph (M*,O*), where M*=in(O*) ∪ out(O*).

Theorem *If a PNS problem is hierarchical then algorithm Hsolve gives an optimal solution of the problem or it finds that the problem has no feasible solution.*

Proof: First we prove that if the algorithm gives a solution then the produced sets (M*,O*) yield a P-graph which is a feasible solution. Since in the determination phase we start to build O* with the set producing P, it follows immediately that P is a subset of in(O*) thus it is contained in M*. Therefore, property (A1) holds.

In a hierarchical PNS problem there is no operating unit producing raw material, thus we get that in (M*,O*) there can be no edge going into a raw material. On the other hand, consider a material X from M*. Then it is an input or an output material of some operating units of O*. If it is an output material then obviously there is an edge going into it in the solution. If it is an input material of some elements of O, then it becomes an element of A during the determination phase. If after that no further iteration comes then X is a raw meterial, otherwise we extend O* with some operating units producing X, thus there will be an edge leading into it in the P-graph (M*,O*). Therefore, we proved that (A2) is valid.

To prove (A3) we have to show that for each operating unit there exists a path from it into a desired product. One can see this statement by induction on the iteration steps of the determination phase. At the beginning we choose such operating units which produce directly desired products. Later in each step we choose operating units which produce input materials of some operating units selected earlier.

Property (A4) is obvious by the definition of M*.

Therefore we can conclude that algorithm Hsolve returns a feasible solution. Now we prove that it finds the optimal one. First we show the following lemma.

Lemma 2 *F(S,j) gives the smallest cost which can be used to produce all materials from set S with j extended costs for each set S which is a subset of M_i for i=0,...,l and for each j=0,....,b..*

Proof of the lemma: We prove this statement by induction on i. For i=0 all elements in M_0 are raw material, therefore, the P-graph which contains only S is a feasible solution with cost 0 for any j. Thus the statement is valid for i=0. Now suppose that i<l and Lemma 2 holds for all subsets of M_i. We will prove it for the subset of M_{i+1}.

Consider now an arbitrary subset S of $M_{i+1,}$ and a feasible solution producing S from the raw materials in the robust model with bound j. In this solution some operating units are selected from level $O_{i+1,}$ denote the set of these operating units by U. Then the input materials of U from level M_i are also produced. If p operating units have extended cost from set U then in(U) has to be produced with minimal cost using j-p extended costs. On the induction assumption this can be done by the cost F(in(U),j-p). Therefore, the cost of producing S will be the maximum of the values F(in(U),j-p)+C(U)+E(U,p). On the other hand, this value was considered in the minimum which calculates F(S,j) thus we obtained that F(S,j) cannot be larger than the cost of producing S. On the other hand, the determination phase produces a solution with cost F(S,j) thus the lemma is valid for the subsets of M_{i+1} as well.

By this lemma we obtain the correctness of the algorithm easily. If F(P,j)<N, then the determination phase builds a solution with cost F(S,j), therefore by Lemma 2 it is optimal. If F(P,j)≥N, then we prove that no feasible solution exists by contradiction. Suppose, we have some feasible solutions. Consider the optimal solution, and denote its cost by OPT. Then OPT cannot be greater than the sum of the extended costs of all operating units, thus OPT<N. But by Lemma 2 we have F(P,b)=OPT and this gives a contradiction.

Theorem 3. *The time complexity of algorithm HSolve is $O(4^k blk)$ which is linear if the parameters k and b are considered as constant.*

Proof. In the initialization part we assign the value of each subsets of R and for each j, thus it takes $O(2^k b)$ time. In each iteration we consider b times all subsets of $M_{i,}$ therefore we perform Step 1,2 of the iteration phase all together at most $O(2^k bl)$ times. In Step 1 in the worst case we have to consider all subsets of O_i, thus we have at most 2^k sets and for each of them the maximum in Step 2 has to be determined which needs at most O(k) time. Therefore the total time needed to perform the iteration part is $O(4^k blk)$. Finally the determination phase needs time O(l).

Remark: We must note that the running time of the algorithm is exponential in the width of the problem, therefore it is effective only for small k-s.

Conclusions

In this work we defined the robust extension of the combinatorial PNS model which can be used in the analysis of process network synthesis and workflow problems. We presented the first results in this area: an exponential time branch and bound based algorithm for the solution of the problem in the general case, and a faster polyniomial time algorithm for the solution of thin hierarhical problems. We think so that there are many further interesting questions to investigate. As far as the branch and bound algorithm is concerned it would be interesting to develope further versions using other bounding functions or other material selection rules, and to compare the efficiency of the different versions. Morever, it is also an interesting question how one can extend the heuristic algorithms presented for the standard PNS problem.

On the other hand, we defined and studied only the combinatorial model in this paper, which can be used to analyse the structure of the networks. It is an interesting question how to extend this model to the quantitative PNS model where the cost of an operating unit is not a constant but depends on the amount of materials used and produced by it. Moreover, in modeling workflow problems it would be important to take into account the shifts, it is also an interesting question to extend the robust PNS model into this direction. Finally, we mention that there are some results on the PNS problem where some other objective besides the cost is also taken into account (see [13]) it would be an interesting question to extend the robust model to handle this situation as well.

Acknowledgement

This work was partially supported by the European Union and the European Social Fund through project Telemedicina (Grant no.: TÁMOP-4.2.2.A-11/1/KONV-2012-0073).

References

[1] Blázsik, Z.; Holló, Cs.; Imreh, B.; Imreh, Cs.; Kovács, Z.: On a Well Solvable Class of the PNS Problem, *Novi Sad Journal of Mathematics*, **30**, 21-30, 2000

[2] Blázsik, Z.; Holló, Cs.; Imreh, Cs.; Kovács, Z.: Heuristics for the PNS Problem, Optimization Theory, Mátrahaza 1999, *Applied Optimization* **59** eds. F. Gianessi, P. Pardalos, T. Rapcsák, Kluwer Academic Publishers, Dordrecht, Boston, London, 1-18, 2001

[3] Bertsimas, D.; Brown, D.; Caramanis, C.: Theory and Applications of Robust Optimization, *SIAM Review*, **53**, 464-501, 2011

[4] Bertsimas, D.; Sim, M.: Robust Discrete Optimization and Network Flows, *Mathematical Programming Series B*, **98**, 49-71, 2003

[5] Fan, L. T.; Kim, Y.; Yun, C.; Park, S. B.; Park, S.; Bertok, B.; Friedler, F.: Design of Optimal and Near-Optimal Enterprise-Wide Supply Networks for Multiple Products in the Process Industry, *Ind. Eng. Chem. Res.*, **48**, 2003-2008, 2009

[6] Friedler, F.; Fan, L. T.; Imreh, B.: Process Network Synthesis: Problem Definition, *Networks*, **28**, 119-124, 1998

[7] Friedler, F.; Tarján, K.; Huang, Y. W.; Fan, L. T.; Graph-Theoretic Approach to Process Synthesis: Axioms and Theorems, *Chem. Eng. Sci.,* **47(8)**, 1973-1988, 1992

[8] Friedler, F.; Varga, J. B.; Fan, L. T.: Decision Mappings: a Tool for Consistent and Complete Decisions in Process Synthesis, *Chem. Eng. Sci***,** **50(11),** 1995, 1755-1768

[9] Holló, Cs.: A Look Ahead Branch-and-Bound Procedure for Solving PNS Problems, PU.M. A., **11(2)**, 2000, 265-279

[10] Imreh, B.; Friedler, F.; Fan, L. T.: An Algorithm for Improving the Bounding Procedure in Solving Process Network Synthesis by a Branch-and-Bound Method *Developments in Global Optimization*, editors: I. M. Bonze, T. Csendes, R. Horst, P. M. Pardalos, Kluwer Academic Publisher, Dordrecht, Boston, London, 1996, 301-348

[11] Imreh, B.; Magyar, G.: Empirical Analysis of Some Procedures for Solving Process Network Synthesis Problem, *Journal of Computing and Information Technology*, **6**, 1998, 372-382

[12] Imreh, Cs.: A New Well-Solvable Class of PNS Problems, *Computing,* **66**, 289-296, 2001

[13] Imreh, Cs.; Kovács, Z.: On Pollution Minimization in the Optimization Models of Process Network Synthesis, Chemical Engineering Transactions, **7(2)**, 565-570, 2005

[14] Monaci, M.; Pferschy, U.: On the Robust Knapsack Problem, *Optimization Online*, http://www.optimization-online.org/DB_FILE/2011/04/3019.pdf, 2011

[15] Süle, Z.; Bertok, B.; Friedler, F.; Fan, L. T.: Optimal Design of Supply Chains by P-Graph Framework Under Uncertainties, Chemical Engineering Transactions, **25**, 453-458, 2011

[16] Tick, J.: Fuzzy Extension to P-Graph-based Workflow Models, Proceedings of the 7th IEEE International Conference on Computational Cybernetics, ICCC 2009, 2009, pp. 109-112

[17] Tick, J.: P-Graph-based Workflow Modeling, Acta Polytechnica Hungarica (ISSN: 1785-8860) 4: (1) pp. 75-88 (2007)

[18] Tick, J.: P-gráf alapú workflow modellezés fuzzy kiterjesztéssel 95 p. 2007 (Disszertáció: PhD)

[19] Tick, J.: Workflow Modeling Based on Process Graph, Proceedings of the 5th Slovakian - Hungarian Joint Symposium on Applied Machine Intelligence and Informatics SAMI 2007, Poprad, Slovakia, January 25-26, 2007, pp. 419-426, ISBN:978-963-7154-56-0

[20] Tick, J., Kovács, Z.: P-Graph-based Workflow Synthesis, Proceedings of the 12th International Conference on Intelligent Engineering Systems, INES 2008, Miami, USA, February 25-29, 2008, pp. 249-253, ISBN: 9781424420834

[21] Tick, J.: Fuzzy Control Systems Based on Parametric T-Norm Function, Proceedings of the 4th International Symposium on Applied Computational Intelligence and Informatics SACI 2007, Timisoara, Romania, May 17-18, 2007, pp. 215-218, ISBN: 142441234X

Evaluating Enterprize Delivery Using the TYPUS Metrics and the KILT Model

George L. Kovács

Computer and Automation Research Institute of the Hungarian Academy of Sciences, Kende u. 13-17, 1111 Budapest, Hungary
University of Pécs and Technical University of Budapest
gkovacs@sztaki.hu

Abstract: The goal of this work is the technical, ecological, environmental and social examination of the life-cycle (LC) of any product (consumable, service, production) using the TYPUS metrics and the KILT model. The life-cycle starts when the idea of a product is born and lasts until complete dismissal through design, implementation and operation, etc. In the first phases requirements' specification, analysis, several design steps (global plan, detailed design, assembly design, etc.) are followed by part manufacturing, assembly, testing, diagnostics and operation, advertisement, service, maintenance, etc. Then finally disassembly and dismissal are coming, but dismissal can be substituted by re-cycling (e.g. melting the metals) or re-use (used parts applications). Qualitative and quantitative evaluations of enterprise results are supported by the new models and metrics.

Keywords: enterprise delivery; environment; life-cycle management; energy; sustainability

1 Introduction

The topic of this paper is the investigation of the life-cycle-management (LCM), or simply life-cycle (LC) of any product (parts, assemblies, services, production units, consumables, etc.) for the total life-cycle. Life-cycle starts when the idea of a product (service, etc.) is born and lasts until the disposal (or reuse, recycling) of the given item. In the first phases there are the requirements' gathering, analyzes, planning (global plans, detailed plans, assembly plans, etc.), then comes parts' manufacturing (including assembly), testing, diagnostics, and the manufacturing is done. Control, service, advertisement and sales are parts of the LCM, as well as maintenance and repair, etc. Finally when the product is worn out or becomes useless there is the destruction (disposal) or reuse (e.g. disassembly of used, but useful parts) or recycling (e.g. melting again). These processes are often modeled and simulated with different models and simulators depending on designer, user, manufacturer, operator, and depending on the complexity of the product. Reuse and recycling were not taken seriously for a long time, now they are understood more and more widely – due to the real or expected energy-, material- and water-shortage in the near and remote future. Even working-force shortage can be imagined in spite of the recently high (more than 10% in Hungary in 2012) unemployment rate in the developed countries.

We plan to investigate the ecological footprint (environmental effects, energy- and material- consumption, CO_2 emission, etc. of manufacturing and usage (operation) of different products and services. The detailed study of these issues and the evaluation of international literature and of measures' of governments and public bodies can be found in [2] and [3], in the books of Prof. Michelini.

A note: in November, 2010 a plenary lecturer at the Hungarian Academy of Sciences at the "Day of Science" stated the following: "A cup of coffee needs the consumption of 600 liter water if everything is taken into account". It means everything from watering the plants to washing the dishes, etc. Another interesting issue can be the electric (or hybrid) car, if all effects of battery acid and heavy metals, etc. are taken into account from mining to dangerous crap.

The KILT model and TYPUS metrics will be explained and used for our study, for details see [1], [2] and [3].

The main goal of our study is to model and quantify the complete delivery (all products, side-products, trash and effects of them, i.e. all results), and changes of the delivery of a firm, and to model all interesting and relevant steps of the LC (or LCM).

In connection with the already well known 'extended enterprises' and 'sustainable development' some notions and their relationships are worth to take a look – without going into details - for the LCM point of view:

- To use extended product – defined based on the philosophy of the extended enterprise
- To distinguish between the product itself and the services provided by the given product
- To distinguish between the tangible and intangible (non-material) aspects of a product
- Fig. 1 needs some explanation. It is a rather simplified view of some main players in the production/service arena, however it still shows quite well certain main qualitative relationships among the given players. We believe that these can be used to understand what is going on in our (engineering-manufacturing-sustainable) world. The TYPUS/KILT metrics, methodology and model give us possibility to better understand and evaluate production results and their components in terms of the defined K, I, L, T notions and values. They give us a method of calculations and comparisons based on realistic values, and not only qualitative, but quantitative evaluations can be done, too. The side effects and 2nd and 3rd order effects, etc. mean the following, explained using a simple and simplified example instead of a precise definition: let us consider the production of a hybrid car.

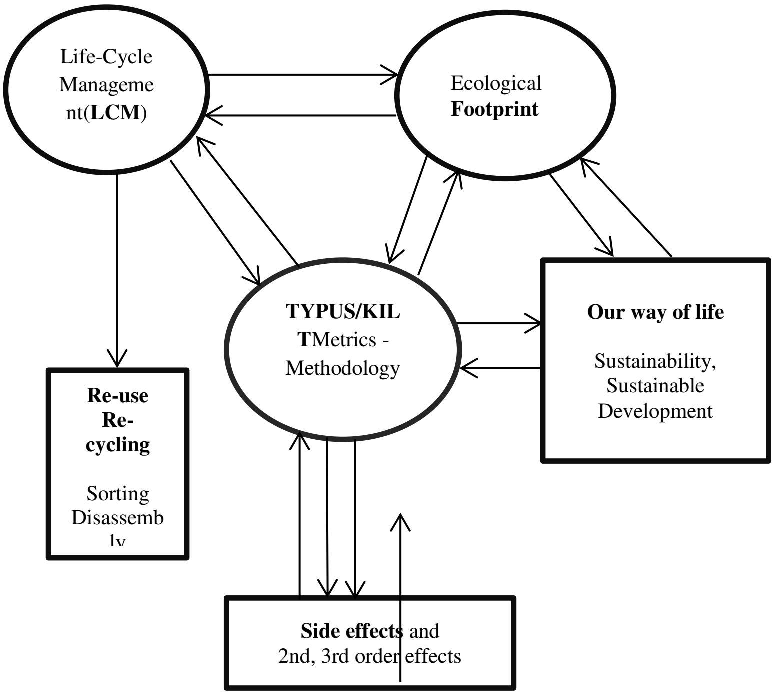

Life-Cycle Manageme nt(LCM)
Ecological Footprint
TYPUS/KILTMetrics - Methodology
Our way of life
Sustainability, Sustainable Development
Re-use Re-cycling
Sorting Disassembly
Side effects and 2nd, 3rd order effects

Figure 1
Some basic LCM-footprint-KILT relationships

It means (today) (among others) to produce and build in two engines, two engines need more metal than one (side effect), to produce more metal we need more electrical energy and more ores (2nd order side effect), to produce more electricity more fuel is necessary and to produce more ores needs more miners' work (3rd order side effect), etc., and it could be continued. It is a hard task to know how deep and how broad we should go with such calculations. And if we take a look at our example there are several other viewpoints (breadth increase) that could be taken into account. Just one example: the increased water consumption during mining. The other blocks of Fig. 1 are closed to trivial, there is no problem with understanding and interpretation.

Our future results will help to examine not only working, operating enterprises (all defined as products), but even enterprises under design can be evaluated taking into account several environmental, ecological, human and other issues of sustainable development or simply of sustainability. This way life-cycle engineering (LCE), or simply LC design can be supported, too. The whole system can be used as a kind of simulation; it gives assistance to imagine something that will work in the future only. To properly understand the need of measurement and evaluation of enterprise results (delivery, all output) in every phase of the LC it is necessary to analyze products and production in details from several viewpoints, including environment effects. This will be done in Chapter 2.

Today, the ‘ecologism’ discovers the damages of the industrial economy, greatly based on the manufacture transformation efficiency, making withdrawals from finite earth stocks, and piling up waste and pollution amounts, exceeding the natural recovery potentiality. Hence, natural opposes to artificial, in view to limit dumping and contamination, moving back the growth rate at pre-industry figures. In the anthropic vision (relating to human beings or the period of their existence on earth), the nature never opposes to the human, being, in reality, only helpful complement of the progress. Some other visions had different nature in the past and even today. However, the term “natural capital” is a recent designation, openly linked to the ecology movements.

The designation is notably effective, when the accounting schemes are requested. As an example, let us take “materials”. The production of everything necessitates materials, compulsorily to be taken from somewhere. The processes require appropriate monitoring. The balances split up to the renewable and the non-renewable resources. The latter shall be classified in terms of (direct, instrumental, etc.) usefulness, (express, subsequent, etc.) toxicity, (local, global, etc.) rarity, and so on, assuming the unidirectional flow from provisioning, to useful products. The point-of-sale denotes the manufacturer’s interest end, leaving entire responsibility of the use, misuse and disposal to the purchasers. These irresponsible supply chains are made legal by the current bylaws, to be, in reality, better classified as populism superficiality or swindle.

The description is coherent with the economic productivity, specified in the manufacturing phase, and exploiting the instant supply/demand balance in the materials’ provision, with no worry for the context. The tax systems follow similar logics: e.g., the VAT moves along the supply chain, charging the increments, utterly neglecting the correlated spoil amounts and effluence levels. This method is faulty: the earth stocks will run out; the environment will turn lethal; only, the today consumers (producers and buyers) make profit.

2 Products, Progress, Life-cycle, Energy, Environment

The already outlined remarks require revising the current attitude about the supply chains. The following – really different, but corresponding – points (levels and ranks) are hereafter could be recalled:

- at the agricultural products' level: aspects in the farm and breeding activity, based on iterating controlled biological cycles
- at the industrial products’ level: facets in the manufacture doing, grounded on better consciousness of the decline threats

- at the business stability rank: the model for the evaluation of all (four) capital assets (K, I, L, T), involved in the shop-floor processes
- at the lawfulness rank: the (example) metrics for the tangibles' appraisal, by standard (legal metrology) bookkeeping.

The above four points stress world-wide examined concepts: the need to give autonomy to the "natural capital", in view that the visibility might motivate enhanced eco-consciousness; the manufacturers' responsibility on the product lifecycle, leaving aside the misleading restriction on the consumers. Both concepts are, perhaps, obvious, but are late in the awareness of many current operators, which prefer naive (actually, egoistic) simplifying assumptions, not to give visibility to the ecology cycle. On these points, more virtuous supply chains are deemed to establish, aiming at better balancing the available resources, not to permit the hoardings and lootings of confident profiteers.

2.1 The Agricultural Products

These were perhaps the first "products" in the history of mankind. The man controlled biological cycles are archaic ideas, discovered in the pre-history times. The domestication of animals and vegetable species is a surprising success of the group selection, leading to the farming communities, leaving behind the starvation risks of the pickers/hunters' assemblies. The human intervention on the earth surface has been, up to now, so invasive, to make hopeless understanding the today earth surroundings are subject to merely spontaneous (with no artificial human action) changes of the evolutionism. We do not discuss this important issue in the recent study in details, as it is out of our scope today.

2.2 History Outlook and Progress in Industry

The man controlled industrial cycles are recent, first fulfilled in some earth regions. The industrial revolution is considered to be typical outcome of the capitalism, especially promoted by the nation-state combined backing of the venture companies, which are entitled to pursue public and private profits. The widespread exploitation of artificial energy is a primary technological innovation; the work-organization efficiency is a fundamental enabling complement. The two facts alone do not show radical divides. The typical effects of the steam power were already acknowledged. The job allotment rules were known practice in the arsenals ship-building. The revolution is a shared issue of enterprises immersed in the new specific cultural surroundings of a subset of populations.

The industrial revolution, as a recent achievement, has throughout descriptions, making possible to assess early developments and cross-links more sharply than for the agricultural revolution. The "industry" is defined as the business establishment, which nicely exploits the structured work-organization and the facility-integration. The meaning is, either: branch of trade or manufacture, assuring productive efficiency; or: diligence and habitual employment in useful activities (industrious is equal to diligent).

The industrial revolution has motivated altering the meaning of several things. The process is somehow mirror of the one pursued by the word "culture", from the land cultivation, to the people instruction. In its original form, the industry has been based on the scientific work-organization and the economy of scale through the mass-production. The recent robot technologies have brought out the intelligent work-organization and the economy of scope through one-of-a-kind manufacture. The industry patterns undergo changes; the variations amplify the opportunities, as it was the case, with the agricultural revolution.

The industrialism has promoted the affluent society and consumerism. The drawbacks are well assessed; they open impending threats to the earth progress. The manufacture process concerns non-renewable resources. As well, the irreversible transformations deteriorate the surroundings, with damages to the bio-sphere. The changes towards the intelligent work-organization depend on the integration of computer engineering tools. These have already been recognized as key help "to de-materialize", and relevant support for the "natural capital" bookkeeping.

As soon as the negative aspects of the industrialism are made clear, the search for remedies shall start. The total suppression of the material goods is a non-sense. The burning up of inanimate stuffs is standard process to carry on the vital cycles. The resort to artificial energy highly (and selectively) speeds up the consumption rates, but again, the simple suppression of the option is meaningless, not to wipe out the current quality of life. The possible remedies are, quite sadly, only partial and temporary:

- to augment the tangibles' productivity, obtaining larger output, while lowering the native exploited input (process effectiveness and recovery/ reclamation closure);
- to discover suited "to re-materialize" cycles, renewing the amount of useful earth stocks, at artificial rates ("robot age" technology transformations).

The success of remedies will not aim at unlimited progress, rather at bounded growth, consistent with the weak anthropic bio-sphere's duration forecasts. The current engineering concerns are analyses on how to improve the resource effectiveness. They include a mix of opportunities, such as the following, for example:

- to reinvent the manufacture cycles, under resource manager liability;
- to avoid waste, planning closed flows, chaining outputs into inputs
- to deliver functions, replacing goods, under unified overseeing;

- to invent domotics (home automation), optimizing the energy controlled delivery
- to supply lifecycle service, fulfilling maintenance and refurbishing
- to perform reverse logistics, up to mandatory recovery targets

The example list show already well understood businesses, which are deemed expanding in the near future. The “to re-materialize” remedy is a longer term issue, involving, most likely, the agricultural ideas, to deal with animated resources, and to exploit suitable bio-mimicry transformations, which enable the related self-reproduction capacities. Now it is impossible to abolish the industrial products, as in the past, these did not remove the land produces. The remedies aim at finding out conservative tracks and replacement means, according to suitably planned restoring/remediation criteria.

The doable remedies have an extraordinary imperative trait of urgency. The climate changes hurry up the need of lowering the contamination, starting by the CO_2 emission. This comes from the oxidization processes, including the ones of animal life. The atmosphere of the earth is highly oxidizing, having the 21% of oxygen. Some 4.5 billion years ago, the atmosphere was highly reducing. The actions of photo-synthesis moved towards the today balance; without it, the CO_2 would become dominant. The current composition is only marginally stable: at higher O_2 concentration, self-combustion establishes (by 24% of O_2). Furthermore, all living beings needs energy, and mainly exploit the $2H_2 + O_2 => 2H_2O$, highly exothermic reaction, which allows reaching to life-suited temperature.

The today atmosphere has about 0.05% CO_2 (78% nitrogen, 1% argon). Bigger CO_2 emissions have rising side falls-off (greenhouse effect, etc.), altering the biosphere equilibrium. The real dynamics depends on multiple factors. Several models are in use, to simulate potential scenarios. The control of the CO_2 emission is a critical request, to preserve the rather peculiar earth habitat, having negligible CO_2, in spite of the highly oxidizing atmosphere. The environment-industry will become key business in the tomorrow world, which adds to the entrepreneurial developments. The innovation is technical capital challenge.

2.2.1 Some More Historical Energy and Environment Hints

It is worthwhile to mention that in the childhood of the author, at the beginning of the nineteen fifties the western countries started to understand the problems of keeping our environment “green” and “clean” taking care of some natural resources (for example: “Keep America Clean !!”). However this understanding was rather partial. The first steps to solve some problems of the fast decrease of easy availability of some metals and of clear water were done, however the whole world believed that energy and fuel/gasoline supply is infinite. At least they were behaving as it were. This led to the huge cars in the USA using too much gasoline and to the huge refrigerators and air-conditioning equipment using huge amount of electrical energy, etc. And we, the whole world were proud of the increasing amount of used energy. Somehow similarly to the recent Moore's low in microelectronics mankind was proudly announcing the fast increase of energy consumption. Moore's law describes a long-term trend in the history of computing hardware whereby the number of transistors that can be placed inexpensively on an integrated circuit doubles approximately every two years. Relative data (increase of energy consumption in %) of the socialist countries, led by the Soviet Union were even higher, as they started rather low.

Nature and environment were different issues in our countries. While the Western hemisphere already understood from the fifties of the 20th Century the rules of co-existence with the environment the Eastern block was still fighting to change the natural environment and to win against nature. Even exaggerations as changing the flow directions of some rivers, irrigation of deserts for agricultural usage and introduction/deployment of several different plants and animals in areas where they never lived before – and cannot live today – were supported and resulted in catastrophic results in most cases.

In [4] we find the following: Electrical-energy production stood at 329 billion kilowatts in 1950 in the USA, 232 per cent more than the 142 billion in 1940, with the cost per kilowatt steadily declining. It means an increase of approx. 23%/year. Most of the electrical energy was produced with power stations run be fossil fuels (natural gas, coal and crude oil). Soon after the end of World War II a vast array of new electrical devices made its way into households, including dishwashers, freezers, dryers, vacuum cleaners, ranges and ovens, and refrigerators. The availability of smaller items such as vacuum cleaners increased through door-to-door sales, and larger items were popular, too.

The next twenty-two years, between 1950 and 1972, total world energy consumption increased 179 per cent, much faster than population growth, resulting in a doubling of per capita energy consumption. This means a decrease in comparison to the previous 10 years, however still an average increase of 8%/year. Oil accounted for much of this increase, rising from 29 per cent of world energy consumption in 1950 to 46 per cent in 1972. By 1973, oil accounted for 47 per cent of U.S. energy consumption. Western Europe and Japan were even more dependent on oil for meeting their energy needs; by 1973 oil accounted for 64 per cent of west European energy consumption and 80 per cent of Japanese energy consumption. For more details see [5]. After the 1972-73 oil crisis the increase of energy consumption was a rather low value, 1%/year only in average until 1992. Governments understood a lot and had several measures worldwide. Then from 1992 to 2000 an increased increase (2.5%/year) meant that the fear of shortages earlier was exaggerated [6].

Today the whole world, all at least most countries understand the importance of natural resources, environment, and based on this understanding reuse and recycling are getting more and more important in everyday life, as well as the decrease of CO_2 emission, etc. These all request to keep energy, water, natural resources, manpower, etc. consumption in a moderate, sustainable level. This leads to sustainable development, or even to sustainability.

2.3 Business and Law Aspects

These are extremely important but out of the recent scope, and they will not be discussed in this paper.

3 Shop-floor Modelling – Industrial Approach

The manufacturing/production activity cannot be suppressed, even if the transformation of raw materials based on artificial energy is a paradigmatic example of consumerism and natural capital decay. When planning for remedies, three facts ought to be taken into account:

- to recognize that the natural capital use requests refunding of all the withdrawals;
- to assess and to bill the materials' costs, with resort to fair legal metrology schemes;
- all factors, players, parts of the LC should be measured and evaluated for all time spans.

3.1 The KILT Model and the TYPUS Metrics

TYPUS metrics means Tangibles Yield per Unit of Service. It is measured in money – on ecological basis. It reflects the total energy and material consumption of (all) extended products of a given unit, e. g. of an enterprise.

The metrics assumes:

- to define a scale for measuring the life-cycle function supplied by the artifact,
- to record the material and energy provisions during the manufacturing phase,
- to record the material and energy provisions during the operation service,
- to evaluate the material and energy recovery at dismissal, reuse and recycling.

To demonstrate the difference between tangibles (for example a piece of metal) and intangibles (defined by the shape and function, for example a spoon made from the metal), there are some simple provisions:

Provision of tangibles: Extended warranties (supply maintenance) and Temporary allocation of artefacts (leased commodities)

Provision of intangibles: Temporary use of artefacts (shared commodities) and Dematerialized assignments (function delivery)

KILT is an arbitrarily chosen implementation of TYPUS, we could imagine other realizations as well. However recently the given definition seems to be the best for the author to be used for the requested goals.

The related TYPUS metrics is further discussed in 3.3.

The refunding needs synthetic models, describing the manufacture processes. With earlier models, the delivered quantities (all outputs), Q, are assumed to depend on the contributed financial (I) and human (L) capitals only.

$$Q = f\,(I, L) \tag{1}$$

Simple relations have been employed, such as:

$$Q^o = \alpha_o xIxL, \ , \ \text{or, incrementally:} \tag{2}$$

$$\Delta Q^* = \beta_o xIxL - \beta_I xI - \beta_L xL. \tag{3}$$

Linear input/output models are assumed for instant marginal description of the quantities, or for related increments around an optimal setting. Enhanced models are in use, to include market entry thresholds (Imin and Lmin), or saturations (Imax and Lmax). Similarly, instead of bi-linear dependence, close to steady state, the lack of symmetry could have resort to modulating exponentials.

The know-how (K) innovation and the tangibles (T) bookkeeping have non negligible effects. The modified – and stronger, more closed to the real life – relationships became:

$$Q = f\,(K, I, L, T) \tag{4}$$

The above given relationships will change to the following:

$$Q^o = \gamma_o xKxIxLxT; \ \ \text{or, incrementally:} \tag{5}$$

$$\Delta Q^* = \delta_o xKxIxLxT - \delta_K xK - \delta_I xI - \delta_L xL - \delta_T xT. \tag{6}$$

Summarizing the different types of capitals and their meaning and some possible components we get:

K: Technical capital – knowledge, technology, know how, etc.- intangibles

I: Financial capital– investment, capital, etc.

L: Human capital– labor, traditional labor, human efforts, welfare charges, etc.

T: Natural capital – tangible resources: material, consumables, ecologic fees, utilities, commodities, etc.

All the contributed technical K, financial I, human L and natural T capitals are included, to supply overt account of the knowledge and tangibles effects. Again, the tetra-linear dependence assumes to operate nearby equilibrium assets. With optimized choices, the negative appendages accomplish the sensitivity analyses, separately accounting the individual capital contributions. K, I, L and T should have their values, while γ_o, δ_o, δ_K, δ_I, δ_L and δ_T are appropriate constants – hard to define and determine them.

The new KILT models reliably describe the delivered product quantities, Q. Lacking one contribution (any of the above constants has a value of 0), the balance is lame, and the reckoned productivity figures, untruthful or meaningless.

The analyses investigate the piling up invariance is against the resort to non-proprietary technologies, or to off-the-market loans, or to work out-sourcing or productive break-up. These models can be modulated with thresholds and nonlinearities.

The tetra-linear dependence means the equivalence of assets alone, and their synergic cumulated action. The company return is optimal, when the (scaled) factors are balanced; the current scaling expresses in money the four capitals (the comparison of non-homogeneous quantities is meaningless; the output Q has proper value, with the four inputs homogeneity). The return vanishes or becomes loss, if one contribution disappears. The loss represents the imbalance between constituent (know-how, money, work out-sourcing, bought semi-finished parts, etc.) flows.

In the bi-linear model, the tangibles (utilities or commodities) are attainable without limits. They do not affect the manufacture business; the affordable growth trend is undefined. The changes in technology, knowledge or know-how, simply, rescale the productivity of tangibles, processed along the material flow. The new model presumes the direct concern of the sustainability bounds for energy saving, pollution avoidance, natural goods preservation, and the likes. These bounds require introducing the figure of the resource efficiency. The replacement of material goods is cost to society, with non-negligible environmental impacts, which shall be paid by the benefits holders (and not poured out on third people and future community).

The scaled T-factor measures the “natural” capital use/misuse, with annexed allegations. The technology K concern has fall-outs at the design, production and sale phases, and, in the manufacture business, is dealt with, for trade fairness, by quality engineering rules. The standard definition of quality, “conformance to specifications”, binds the design (technical specifications files) to the delivery approval testing. Anyway, the technology K comes as primary transformation factor, affecting the manufacture process throughput.

The earlier dependence on the financial I and human L capitals is result, perhaps, of the dialectic opposition between plant owners and labor. The neglect of the tangibles T factor is surprising, as the manufacture duty has no output, without materials and energy. The addition of the intangibles K is characteristics of the "knowledge" paradigms, and an earlier entry could have been devised from the "new economy" deployment. With the manufacturers' lifecycle responsibility, the explicit role of K and T needs to emerge, since the companies' profitability will become critically dependent on how these factors are balanced. The T vs. K dialectic opposition is a challenge for tomorrow.

The productivity bookkeeping is merged into global Q assessments, to provide synthetic pictures, with visibility of the four capital assets, as the enterprise's function/facility means. The manufacture shop-floors generate value chains, providing Q guesses based on the market requests, including the welfarism charges (L factor), according to the enacted rules, and, from now on, the ecologic fees (T factor), following, e.g., the in-progress EU directives. With the KILT models, the four manufacture assets easily apply, without modifying many assessed traditional habits.

The classic micro-economic description, basically, leads to the delivery figure Q^o, maximally concerned by the steady production flow on convenient strategic spans. The flexible automation (with integrated design, in progress, open to time-varying scopes and externalities) leads to ΔQ^*, which corresponds to the incremental delivered product quantities on the tactical spans. The steady description states that the changes in technology, finance, labor and supply can freely be performed, without affecting the throughput (products and services). With the "old" bi-linear model; the tangibles (raw materials, utilities or commodities) are attainable without limits, thus, they do not affect the manufacture business, and the affordable development trend is undefined; moreover, the changes in technology, knowledge or know-how simply rescale the productivity of tangibles, processed along the material flow. The new model presumes direct concern of the sustainability bounds for energy saving, pollution avoidance, natural goods preservation, and the likes; these bounds lead to introduce the concept of resource efficiency. The replacement of material goods is a cost to society, with non-negligible environmental impacts that shall be endorsed by the benefits holders (and not poured out on the community). The fact is dealt with by the scaled T-factor, with serious implications. The explicit concern of the technology K has different falls-out at the design, manufacture and sale phases, and, in the manufacturing business, it is dealt with, as for trade fairness, by quality engineering rules.

Actually, "quality" has the standard definition: "conformance to specifications", binding the design (technical specifications files) with the delivery approval testing; quality functions deployment or similar company-wide quality-control set-ups are later recognized as contrivances to improve efficiency, through zero-defects production. It is updated as "fitness for purposes", with focus on the users' requests. The last definition is ambiguous, unless the quality is measured according to quantitative methods, within the legal metrology frames. This leads to "certified quality", with assessments established by third parties (as compared to sellers and buyers). At the same time, the technology appears as primary transformation factor, affecting the throughput of the industrial organizations. The matter can be looked as well, saying that "quality" is technically specified at the design (conformance to specifications) or at the marketing (fitness for purposes) steps, if the reference standards are defined. On this line, the "quality" is widened to include the eco-consistency demands, and the "certified quality" has to be assessed with resort to the TYPUS metrics, or other equivalent enacted standard. Then, quantitative performance functions are stated, to logically connect effectiveness up-grading and capital investments, for detailing impacts and returns of the (four) fixed assets. By that way, quality engineering becomes the meaningful go-between for company-wide information set-ups (measured with formally specified standards). At this point, to assess the value build-up of actually delivered artefacts, the embedded technology has to be accounted for, as primary production factor (not to be merged in the I-factor, as non-distinguishable acquisition, or in the L-factor, as operators' inherent property).

From our point of view equation (3) and (6) are the most interesting. The resulting ΔQ^* (simply delta Q) can be understood and evaluated as the change of delivery, i.e. the change of value of the product between two steps (phases) of the life-cycle. This way LCM has a new meaning, where the value changes can be followed between any steps, which may be consecutive as well. In this type of understanding we do not have to speak about delivery, but about phases of the PLCM process, with defining the value changes. For example the value difference between a painted car body part and the same part unpainted can be estimated by knowing the added quantity of paint (T), the knowledge, how to do it (K), the amount of money to run the painting shop (I), and the human efforts needed (L). We hope to have genuine results – understanding the value chain – if we go ahead this way.

3.2 Tangibles' Productivity

The environmental protection is man's right at universal range, with outcomes to safeguard the future generations, not today, represented by efficient political parties or governmental agencies. Then, the democratic consensus or international agreements results deprived of justifications, whether limited to place the interests of the today citizens before. The fair socio-political approach, put forward by the "knowledge" paradigms, compels protecting the generations to come, by compensation ways, such as:

- to create a tax system, which consolidates the wealth corresponding to all withdrawal accomplished from the natural capital, following deposit/refund-like arrangements;
- to forbid natural capital withdrawals that exceed quotas, roughly equal to the reverse logistics recovery, or (hopefully) to the bio-mimicry stimulated generation, in view to keep the original natural capital level, by neutral yield.

The first way is formal, since, transforming different capitals, the equivalence criteria are, at least, ambiguous. The second, if coherently applied, faces decay limitations, and today runs into the life quality decrease, towards the thrifty society. It is, moreover, possible to merge the two ways, using the "deposit-refund" choice as first instance, thereafter keenly researching innovative technologies, out of reverse logistics, which perform active replacing resources and full eco-remediation, to achieve neutral yield of the inherited natural capital.

Many unanswered questions exist (e.g., in terms of comprehensiveness). Still, the EU environmental policy looks aiming at united way. The bookkeeping of the tangibles' decrease and pollutants' increase becomes primary demand, with, as side request, the assessment of the restoring onerousness. Then, the closed-cycle economic/ecologic processes are prerequisite of the manufacture markets to come. The analyses need to be quantitative, to make meaningful comparisons, fulfilling the assessments by recognized standards. The consistent closed-cycle appraisal brings to concepts such as the below new metrics (or similar equivalent standards).

3.3 The TYPUS Metrics

TYPUS, tangibles yield per unit service: the measurement plot covers the materials supply chain, from procurement, to recovery, so that every enjoyed product-service has associated eco-figures, assembling the resources consumption and the induced falls-off requiring remediation. The results are expressed in money, resorting to the arbitrariness of establishing stock-replacing prospects. The point is left open, but, it needs to be detailed, to provide quantitative (legal metrology driven) assessment of the "deposit-refund" balance.

The metrics is an effective standard, aiming at the natural capital intensive exploitation. The supply chain lifecycle visibility needs monitoring and recording the joint economic/ecologic issues, giving quantitative assessment of all input/output materials and energy flows. The new tax system has to operate on these data, establishing consumption rates at the input, and pollution rates at the output, to obtain the 'wealth equivalent' of the overall impact (as for the first mentioned way).

When a metrics, such as TYPUS, is adopted, conservative behaviors are quickly fostered. The ecologic bent of the taxing systems becomes enabling spur, to turn the "knowledge" paradigms towards environmental friendly goals. The TYPUS, tangibles yield per unit service, metrics can, of course, take other forms. The objective is to look after capital conservative arrangements, notably, as for the natural assets. In different words, the objective is saving the wealth (the capital), and to tax the consumption (the imbalance of the natural resources).

Today, the eco-fee evaluation is quite obscure, due to political biases. It leads to taxing schemas, which draw from the whole capital, more than in proportion to the actual impact (net consumption combined to pollution). The eco-protection, switched into the individual consumers' business, is starting point, to look for higher efficiency, over the whole supply chain, from provisioning to recovery, using all-comprehensive effectiveness criteria, singly dealing with each capital asset (tangibles' productivity included).

The natural capital bookkeeping will become standard routine of the knowledge society. Today, the economic accounting to detect unlawful habits (crime, repression, etc.) is obvious practice, affecting personal liability. Similarly, the thought-out ecological accounting needs to develop into steady rehearsal, in view to charge the actual consumers (to the advantage of third people and future generations). With the «knowledge» paradigms, the tax regulation restructuring is required, because the socio-political aspects management becomes important contribution to effectiveness. The biggest question is how to distinguish the community's, from the individual's duties.

The solution offered by the capitalism (in western world style) is noteworthy issue. The personal liability is consequence of the independent freedom to organize the exclusive fortune. In the place of shared figures, the individual accountability offers rewards through competition. The averaged taxation of input/output materials and energy flows is simple, being linked to nominal parameters, limiting the control to out-of-all quantities (provisioning and land-filling). This is local communities business, with visible fees refund, depending on the efficiency of averaged balances. The ecological accounting through the TYPUS metrics is more tangled affair, needing to address each single supply chain. The competition shifts at that level, and the lifecycle manufacturers' responsibility is viable bookkeeping charge (under registered overseeing). At least, this is fair practice, deserving attention.

Conclusions and Further Plans

Our real goal is to give some means and tools to calculate different values which correspond to different phases of the life-cycle (LC) of a product. We specially emphasize re-use and re-cycling as important LC phases, due to approaching water-, energy- and raw material-shortages.

On product we mean anything what is used by simple users (a car, a cup, a bike, or a part of them, etc.), or what are used by dedicated users to produce or manage other products (a machine tool, a robot, a house, a test environment, etc.), or which are used to manage everything else (a firm, a factory, a ministry, etc.). We differentiate between simple products and extended products (as traditional and extended enterprise) and between tangible and intangible parts (aspects) and service is taken into account as a product, too.

The TYPUS metrics and the KILT model say, that the delivery of a firm can be calculated with a seemingly simple multiplication of 4 main factors, as Knowledge (innovation) /K/, Investment /I/, Labor (financial and human capital)/L/ and

Tangibles (materials)/T/ – modified by appropriate constants and additional factors. Delivery means the goal-products and all side effects (water consumption, CO_2 disclosure, etc.) together.

Based on these calculations real data can be given on all effects, side effects and 2^{nd}, 3^{rd} order side effects of products and productions, e.g. CO_2 emission can be properly evaluated. As the first practical results we hope to get useful calculations using equation (6) – or a mutation of it - for getting delta Q, what we consider as value change between any two phases of the LCM of any product.

We know that the above given forms cannot yet be used for economically useful calculations, they contain only several ideas and qualitative relationships to go on a right way. We plan to find proper relationships to use our ideas and formulae for real world situations to assist not only designers and engineers in their work, but politicians and other decision makers as well. These studies and their resulting calculations, values and suggestions how to proceed will be in a following study.

References

[1] Michelini, R. C., Kovács, G. L.: Integrated Design for Sustainability: Intelligence for Eco-Consistent Products and Services, in: EBS REVIEW, ISSN 1406-0264, Innovation, Knowledge, Marketing and Ethics, Winter 2002/2003, pp. 81-94

[2] Rinaldo R. Michelini: Knowledge Enterpreneurship and Sustainable Growth, (book, 325 pages), NOVA Science Publ., 2008

[3] Rinaldo R. Michelini: Knowledge Society Engineering, A Sustainable Growth Pledge, (book, 350 pages), NOVA Science Publ., 2010

[4] Gale Cengage: American Decades, ©2000. Published/Released: December 2000, ISBN 13: 9780787650766, ISBN 10: 0787650765, DDC: 973.91, Product number: 172108, Shipping Weight: 49.00 lbs , Price: US $1495.00

[5] Richard Schmalensee, Thomas M. Stoker, and Ruth A. Judson: World Energy Consumption and Carbon Dioxide Emissions: 1950 Ñ 2050, MIT, 1995 study, pp. 1-40

[6] John c. Dernbach, editor: Stumbling Towards Sustainability, book, Published in July 2002 by the Environmental Law Institute, Washington D.C., ISBN: 1-58576-036-6, pp. 1-968

Neural Network-based Indoor Localization in WSN Environments

Laslo Gogolak[1], Szilveszter Pletl[2,3], Dragan Kukolj[3]

[1] Subotica Tech, Department of Automation, Subotica, Serbia

[2] University of Szeged, Department of Informatics, Szeged, Hungary

[3] Faculty of Technical Science, Univ. of Novi Sad, Novi Sad, Serbia
gogolak@vts.su.ac.rs, pletl@inf.u-szeged.hu, dragan.kukolj@rt-rk.com

Abstract: With the advancement of wireless technology even more wireless sensor network (WSN) applications are gaining ground. Their field of application is increasingly widening. This paper examines the WSN application which allows indoor localization based on the Fingerprint (FP) method. The communication between the modules was monitored during the experiment whereby the received radio signal strength indicator (RSSI) values from 5 modules were recorded by a mobile sensor. The received data was used for training of the feed-forward type of neural network. Through use of the trained neural network and the measured RSSI values an indoor localization was realized in a real environment. The neural network-based localization method is analyzed applying the cumulative distribution function (CDF). For the reference model the well-known weighted k-nearest neighbour (WkNN) method was used.

Keywords: Fingerprint localization; WSN; Received Signal Strength; Neural Network; Mobile sensor

1 Introduction

Most technical solutions would not be available without localization processes, even transport is unimaginable nowadays without a navigation device. GPS localization devices can achieve up to 2 cm accuracy in localization for each spot in the world. For GPS localization one has to be in an open space and hold the appropriate device. There are various localization methods, but actually the GPS systems offer the simplest, cheapest and the most accurate outdoor localization technologies. In fact, indoor localization presents the bigger challenge. Although there are numerous applications and methods for indoor localization, none of those are appropriate and accurate enough. Applying different technologies, such as RF, optical, infrared, ultrasound, determination of positions is possible to certain accuracy [16, 17]. Information about an indoor position mostly can be obtained by processing an RF signal sent from known positions. This localization methodology is called the network fingerprinting method [1, 2, 7 and 19]. In this work an indoor localization method is presented using Wireless Sensor Network – commonly known as WSN. There are many solutions for the localization algorithms. In case of the solution offered by Kannan, Guoqiang Ma and Vucetic [15], simulated annealing algorithm was used for the localization. They use the measured distance between the fixed sensor mote (anchors) and the mobile mote for the localization.

In this work the fingerprint method of indoor localization using feed-forward neural network is presented. The WSN sensor modules were placed on fixed positions (anchor motes) in an experimental room. The room is a furnished classroom with computers and glass windows, which is not an ideal environment in terms of the spread of RF signals. A mobile sensor module is part of the system, and with the help of this mobile sensor the RSSI (Received Signal Strength) values between this sensor and the anchor motes can be measured. The RSSI measurements are done in predefined positions with 0.6 m resolution. The received data are stored in a database. This set of data is used to find some useful correlation between the alteration of RSSI values and the position of the mobile mote. This work shows a solution for indoor localization based on trained neural network and created database RSSI measurements. The weighted k-nearest neighbor (WkNN) method is used for the results comparison [18]. While the recording of the RSSI values is done in real environment, the localization is done by software implemented neural network.

2 Wireless Sensor Network

The current trends make the wireless techniques very popular in various fields of application. This low cost technology is available for any purpose. Thanks to the modulation techniques, the quality of the RF transmission is very high. Nowadays the wireless sensor networks are successfully used for, for instance, forest fire detection, monitoring of numerous agricultural microclimates [3, 6] and traffic systems [9].

2.1 The Crossbow Iris Sensor Mote

For the experimental purposes, the wireless sensor motes with weak resources were needed, therefore the Crossbow‘s IRIS wireless sensor motes were chosen. These sensor modules are compliant with the IEEE 802.15.4 standard called ZigBee. The IRIS sensor modules are used for measuring the RSSI values during the experiment. The basic elements of the sensor motes are the single 8 bit low power Atmel’s microcontroller and the ZigBee protocol based wireless module. These microprocessors are from the Atmel’s series 1281 which are working at 8

MHz. The processor contains a 128 Kbytes program flash memory, an 8 Kbytes RAM, a 4 Kbytes configuration EEPROM, and some communication interfaces (UART, ADC, digital I/O, I2C, SPI), whose performance classifies it as a microcontroller with thin capabilities. The third part of the system is the logger flash which is a 512 Kbytes memory, which stores the measurement data. The communication between the sensor modules is done by using the above mentioned 2.4 GHz RF modules. The ZigBee standard communication range in the outdoor environment is up to 500 *m*, but in an indoor environment it is up to 50 *m.* The Iris sensor motes are supported by the open source TinyOS operating system which is programmed in the NesC programming language. The most significant characteristic of the Iris wireless sensor systems with TinyOS is that they can work in the network mode. They can communicate and they can be programmed through the self-organizing ad-hoc Xmesh network. For this work the authors used the above mentioned platform during an experimental setup.

2.2 Measuring the RSSI Values

The radio frequency transceiver is the most important component. It is a high performance RF-CMOS 2.4[GHz] radio transceiver targeted at IEEE 802.15.4 applications. This transceiver chip contains some special IEEE 802.15.4-2003 hardware support such as RSSI computation. The RSSI value is a 5-bit value which is updated every 2 µs. It indicates the received signal power in the selected channel in steps of 3 dB. This value can be in the range of 0-28. The RSSI value of 0 indicates a radio frequency input power of less than -91dBm.

For an RSSI value in the range of 1 to 28, the radio frequency input power can be calculated by following equation:

$$P_{RF} = \mathrm{RSSI\_BASE\_VAL} + 3 \cdot (\mathrm{RSSI} - 1), \tag{1}$$

where the P_{RF} is the radio frequency input power, the RSSI_BASE_VAL is equal to -91 dBm.

2.3 The Measuring Software

2.3.1 The Mobile and the Anchor Node Software

The mobile node and the anchor nodes have the same software for technical reasons. In case of receiving a control message, the node starts measuring the RSSI values. This is done by the following steps: the node receives a message which contains a command to perform *n* measurements with a node that has an ID of *x*. It first sends a query message to the x node, waits for the answer, collects the results and broadcasts it, so that the mote sending the control message can receive it. The mobile node repeats it *n* times. Furthermore, when an anchor node receives a query message, it performs a measure on the incoming message, then sends it back. The mobile node operates in the same way upon receiving this message. In the experiment, the value of *n* was set to 100. The measurement algorithm is shown in Figure 1.

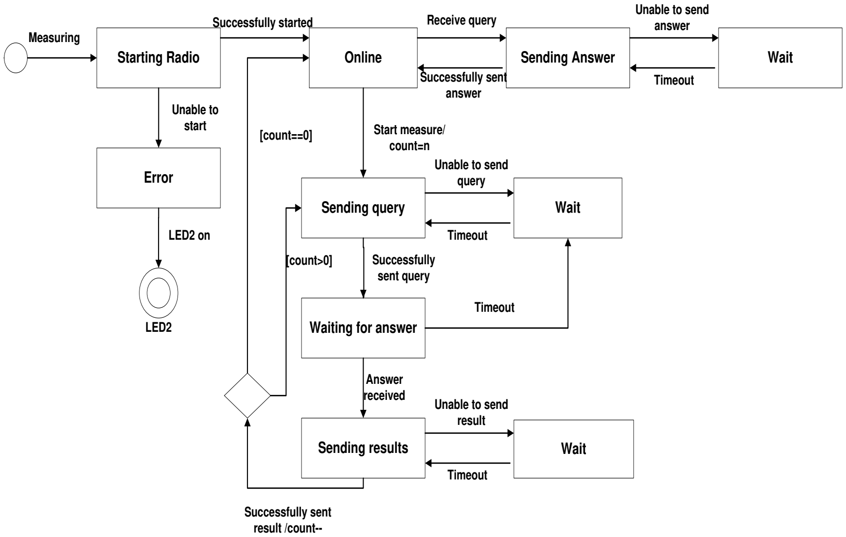


Figure 1
Block diagram of the measuring algorithm

2.3.2 The Base Station Node Software

The software of the base station is very simple, it works only as a gateway to the computer. The user has GUI on the control computer with two text inputs for the location and a button to start the measurement. This program can be manually configured as to which nodes to test (values of *x*) and how many times (value of *n*) along with the timeout. When pressing the button, the program periodically sends messages to the mobile node (through the base station) and waits for the results. After the timeout had been reached, it sends the next message. This prevents the deadlock of the system. In the experiment, the nodes IDs 1, 2, 3, 4 and 5, were tested 100 times, with a 3 seconds timeout. However, there were positions where the signal of an anchor node was too weak.

2.3.3 The Package Format

Four different types of messages are used by the system, all of them based on the default radio message format of the TinyOS system. This framework adds a header to the messages (containing the sender and the target), the length of the message and other metadata. It also handles the *acknowledge* messages. The four user-defined messages are the *control* message, the *query* message, the *answer* message, and the *result* message. The format of four user-defined messages is shown in Table 1.

Table 1
The format of the message packet

Byte	**1**	**2**	**3**	**4**	**5**	**6**
controlMsg	Source		target		count	
queryMsg	Id					
answerMsg	Id	Sig				
measureMsg	Time				signal1	signal2

The *control* message is designed to be the system message. Currently, the only role it has is to start the measurement. It is sent to the mobile node by the anchor node. The *query* message is sent by the mobile node to an anchor. It only contains one unsigned byte, the ID of the message (id) to identify each conversation. As a reply, the anchor sends an *answer* message which contains the message ID (id) and another unsigned byte for the measure (sig). The anchor node calculates this value when it receives the *query* message using a built-in RSSI measure module which provides the RSSI value of each incoming package. The mobile node generates a result message upon receiving the *answer* message. The *answer* message contains a 32 bit timestamp (time, the local time of the node, only for test purposes), the signal strength received from the anchor node (signal1) and the signal strength of the *answer* message (signal2) determined in the same manner as the anchor node. The output of the measuring process is an XML file which contains the count of the measurements and values (the “count” attributes, typically 500), and every measurement has the following format:

- “time” is the time of arrival of the radio packet to the mobile node
- “mote1” is the mobile node
- “mote2” is the anchor node
- “signal1” is the RSSI measured at the mobile node
- “signal2” is the RSSI measured at the anchor node

The typical size was about 69 KB per measurement. The identifiers of the XML documents depend on the measurements due to the simplification of the measurement database.

3 Experimental Measuring and Database Preparation

In the present work the authors use neural network for localization purposes. As in all neural network training algorithms, this application also requires a database for training and testing of the network [4, 5]. The database in this particular experiment is prepared by the measuring of the RSSI values for each $r_j = [x_j\ y_j]$ coordinate by the sensor motes. The location vector $[\mathrm{x_j}\ \mathrm{y_j}]$ denotes the position of the jth reference point. The experimental room where the measurements were done is a research laboratory. The room contains standard laboratory furniture and equipment such as tables, chairs and computers. The structure of the room is a very important factor in case of indoor localization [8]. The radio signals are very sensitive to reflections. The number of windows in the classroom affects the accuracy of the localization. In this experiment the aim was the realization of the fingerprint localization in a non-structured real environment with iron grid-protected windows and with all the interfering factors of the room.

3.1 The Process of the Experiment

The experimental room structure consists of 24 x 10 measurement points (Figure 2 - black points) where the RSSI values for the fingerprint are measured. In Figure 2 all types of points are presented. The distance between the points, the grid resolution is 0.6 m. In the grid system there are 5 sensor motes, and anchors (see the red points in Figure 2), which are in fixed positions. The third type of positions (see the green points in Figure 2) are the unknown positions which are between the measurements points. They will be used for testing the accuracy of the positioning algorithm.

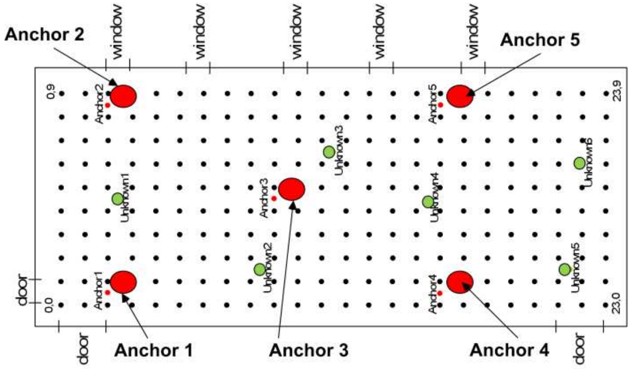


Figure 2
Layout of the experimental setup

During the experiment one measurement is done at every 24x10 (240) position using a mobile sensor mote. The RSSI values are measured between the mobile mote and between the anchor motes. The measurements are done in two directions. The mobile mote measures the RSSI value of the anchor sensor mote and vice versa. In order to obtain satisfactory accuracy, measurement with one anchor mote is repeated 100 times. In the presented experiment there are 5 anchors. For all positions we have 5 x 100 measurements in both directions. Therefore, at the end of the recording, the authors had 12000 measured values for the mobile-anchor links and 12000 measured values for the anchor – mobile links. All of the data are stored in the centralized database for further analysis.

3.2 Data Preparation

The calculation of RSSI values was described in Section 2.2. Due to performance of the hardware platform the highest measured value was in interval from 4 to 16 units. This means the resolution of the RSSI values was only 12 units. This resolution defines the accuracy of the localization results. Unfortunately, the RSSI values can only be measured at resolution dictated by the Iris sensor module. The arrangement of the experimental room is not ideal for the wireless signal distribution. The typical distribution of the RSSI values is shown in Figure 3 for the first 4 anchor motes. From the sample of 100 measurements there were cases when few were invalid. This error occurs because the RSSI values cannot be measured properly. The invalid RSSI values are replaced by the mean values of the neighbouring grid point values. For all positions from the 100 measurements an additional database is created. This database consists of the mean value, median value and standard deviation for every 24 x 10 position and for all the 5 anchors. In fact, this database is used for creating the neural network, whereby 50% of the values are used for training the neural network, 25% of the values are used for the validation and 25% for testing the accuracy of the neural network.

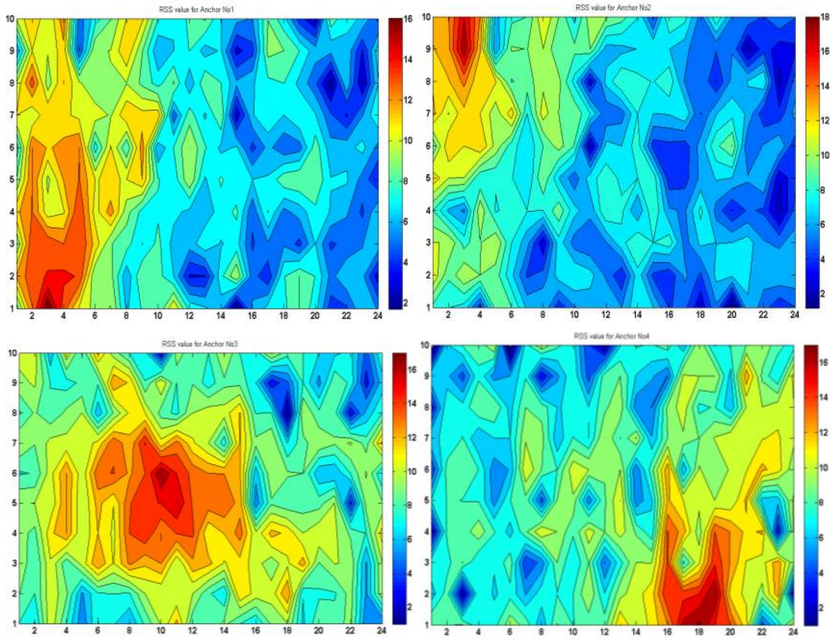


Figure 3

The map of RSSI values for the four anchor motes

4 Weighted k – Nearest Neighbor Algorithm

The weighted *k*-nearest neighbour algorithm is widely known clustering method [10] and can be efficiently applied for fingerprint localization. Assuming that the radio map database of the reference points of RSSI vectors $S \in R^m$ exist, and $S_t \in R^n$ measurements linked to the tracked node are performed, one may indicate the similarity in signal strength vectors of the j^{th} reference point $r_j = [x_j, y_j]$ and the wireless sensor with unknown location. The Euclidian distance p_j is defined as

$$p_j = \sqrt{\sum_{i=1}^{n} (s_{ti}\text{-}s_{ji})^2} \tag{2}$$

determines the similarity between t^{th} current and j^{th} reference signal strength vectors, and If the j^{th} reference point is closer to the node, the similarity value p_j is smaller. That is to say, it may be assumed that the closest reference point would have the most similar signal strength vector to the signal strength vector of the tracked node. The algorithm has a parameter, *k* which affects the accuracy of the method. This parameter also needs to be determined. In this case *k* refers to the number of reference points. The following step of the algorithm is finding the *k* nearest reference points $v_j$with the smallest Euclidean distances p_j, *j=1, k.* A weighting factor w_j, *j=1,k* is associated with the *k* reference points with the most similar RSSI vectors to the RSSI S_t measured at the tracked wireless sensor, based on their *k* smallest calculated Euclidean distances from that node, and calculated according to Equation (2). The weighting factors are inversely proportional to the square of the Euclidean distance, as shown by the Equation (3).

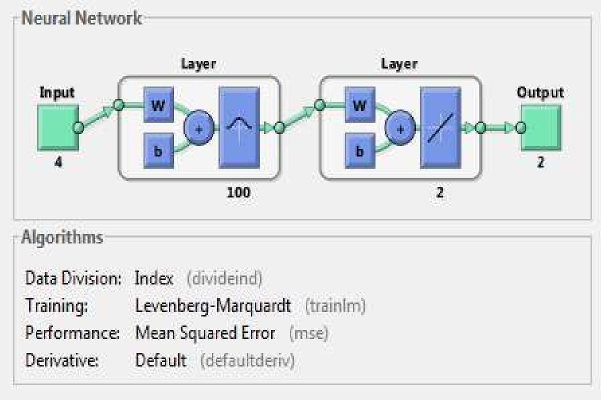
$$w_j = \frac{{}^{1}\!/_{p_j^2}}{\sum_{i=1}^{k} {}^{1}\!/_{p_i^2}} \tag{3}$$

The unknown coordinates of the tracked wireless sensor are estimated by the equation $r_x = \sum_{i=1}^{k} w_i \, r_i$. Finally the estimation of the coordinates is completed. Consequently, this method is called the weighted *k*-nearest neighbor algorithm (WkNN).

5 Neural Network for Position Determination

This section presents a neural network clasterization method for determination of the position of mobile sensor node. This method has been chosen because in this case the recorded RSSI database can be used for supervised learning of the neural network [11, 12]. The type of the neural network is feed-forward. The implementation of the Levenberg-Marquardt (LM) training algorithm, which is chosen for the training of the neural network, was written in Matlab software. The

Levemberg-Marquardt training algorithm is the fastest and most efficient training method, but it requires a significant amount of working memory [13, 14]. In case of applying feed-forward three-layered neural networks, the number of neurons used in the hidden layer and also the type of the activation functions used in the neurons are both significant free parameters. During the simulations, the best results were achieved by using the "tansig" type of activation functions in the hidden layer, and "purelin" type in the input and output layers. The chosen structure of the used neural network is presented in Figure. 4.



Algorithms

Data Division: Index (divideind)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Derivative: Default (defaultderiv)

Figure 4

The Structure of the neural network in case for the four inputs (RSSI values from four anchor)

The neural network has 5 inputs and 2 outputs. The measured RSSI values from the 5 anchors are the input, while the outputs, the *x* and *y* respectively denote the coordinates. The structure of the matrix of the RSSI values is the following:

$$S = \begin{bmatrix} S_{11} & S_{1i} & \cdots & S_{15} \\ S_{j1} & S_{ji} & \dots & S_{j5} \\ \vdots & \vdots & \ddots & \vdots \\ S_{N1} & S_{Ni} & \cdots & S_{N5} \end{bmatrix} \tag{4}$$

where s_{ji} denotes the RSSI perceived i^{th}anchor, $i \in (1,5)$ at the j^{th} reference point. The output coordinates are used in the matrix form shown below:

$$r = \begin{bmatrix} x_1 & y_1 \\ x_j & y_j \\ \vdots & \vdots \\ x_N & y_N \end{bmatrix} \tag{5}$$

As it has been mentioned previously, the second database is used for the neural network training, where the mean, median and standard deviation values are listed. There are some other measurements apart from this database, whose coordinates are beyond the grid resolution. These points are the real test samples, for these coordinates the neural network provides interpolated values as there are no RSSI values for the training. In the process of training and testing the neural network it can be seen that the accuracy and the speed of the training process depends on the numbers of hidden neurons. The following section demonstrates dependence of the accuracy of the suggested neural network based on the number of the hidden neurons.

6 Test Results

The results of the experiments using the neural network trained by the LM method are compared to the results given by the weighted *k* nearest neighbor (WkNN) algorithm. To represent the similarity between the estimated position of the tracked wireless sensor and its true coordinates, the Euclidean distance error was used as the basic performance metric, which represents the distance between the estimated position of the tracked wireless sensor and its true coordinate. Furthermore, the cumulative distribution function (CDF) of the distance error of all location estimates for all measurements fully describes the estimation characteristics.

During the process of testing, the “unknown points” mentioned in Section 3.1 were used for the accuracy of localization evaluation (see Figure 2 – green points). Since these points are positioned between the grid, their RSSI values are not found in the localization database.

Firstly, the localization accuracy of the neural network and WkNN method is compared. The different figures show the accuracy of the localization depending on how many anchor motes and their RSSI values were used during the localization process. Figure 5 shows the localization accuracy of the designed neural network and the WkNN methods where the RSSI values of the all five anchor motes have been used.

Figure 6 shows the 3 different structures of both methods. In the case of the neural network, HN100, HN200, HN300 refer to the number of 100, 200 and 300 hidden neurons, respectively. In the case of the WkNN method the CDF function is shown depending on the number of reference points (k=1,3,5).

As it can be seen, when all the 5 anchor motes were applied, all the versions of the WkNN provided a more accurate localization. The WkNN achieved the highest accuracy when 4 anchors were applied in the localization process. These anchor motes were placed in the four corners of the experimental room. These were the anchors 1, 2, 4, and 5. This case is shown in Figure 6.

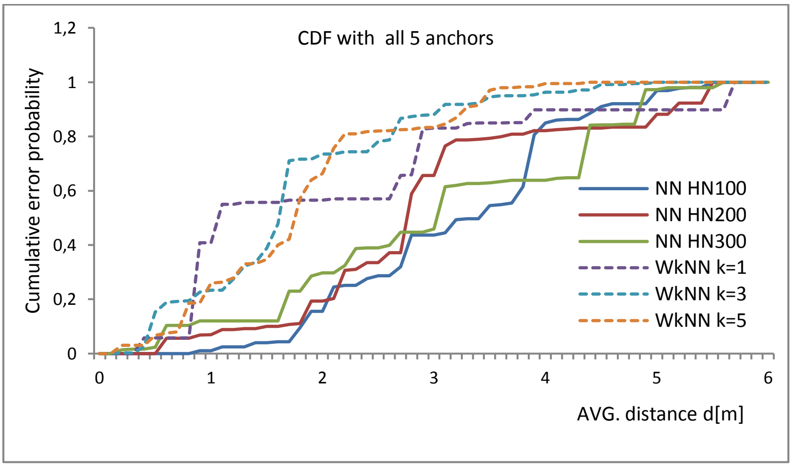


Figure 5
Cumulative distribution functions for distances in case of 5 anchors

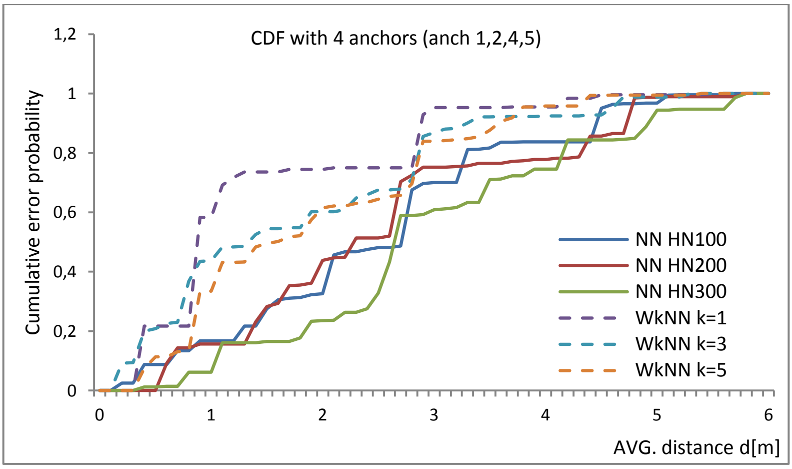


Figure 6
Cumulative distribution functions for distances in case of 4 anchors

Figure 6 shows the increase of the accuracy of the neural network too, whereby the number of neurons has not much of an effect on the accuracy. However, the highest accuracy in localization was achieved by the WkNN method with k=1.

Figure 7 shows the CDF function when only 3 anchor motes were used for localization. These sensor motes are anchors 1, 2, and 5. These are the motes which divide the experimental room diagonally. In this case the accuracy of the localization with the WkNN method significantly decreases. Consequently the neural network has an advantage in this case, because it can provide a more accurate localization than the WkNN method.

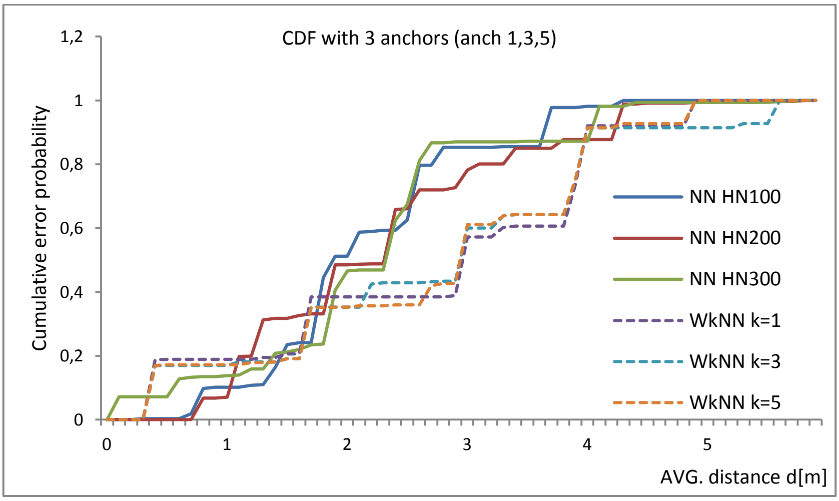


Figure 7
Cumulative distribution functions for distances in case of 3 anchors

The next two figures show the accuracy of the localization depending on the number of anchor motes in the case of the WkNN with k=3 (Figure 8) as well as in the case of the neural network with the structure of 100 HN (Figure 9).

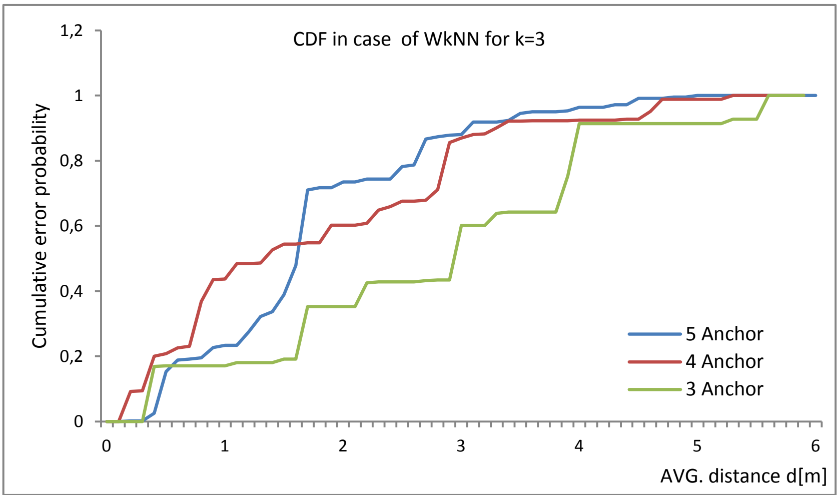


Figure 8
Cumulative distribution functions for distances in case of WkNN for k=3

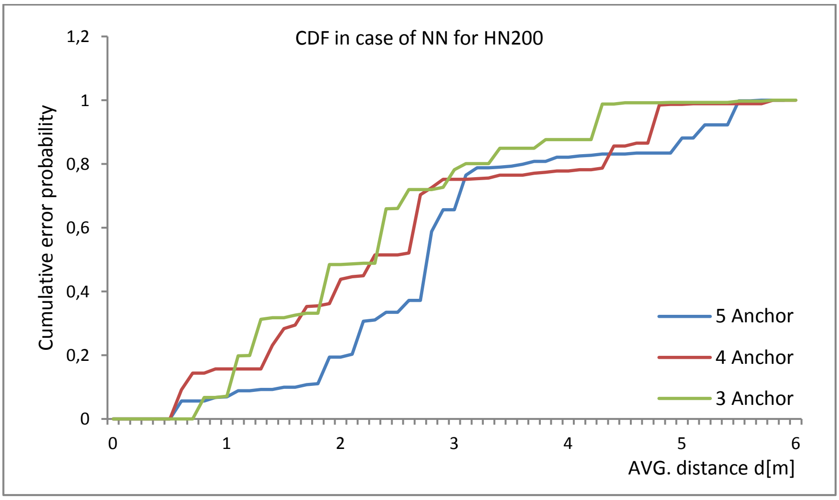


Figure 9

Cumulative distribution functions for distances in case of WkNN for k=3

The localization process with the neural network has shown higher accuracy when less anchors were used. This is due to the greater robustness of the NN on increased uncertainty in conditions with less information. Additionally, the position structure of the anchors can also affect the accuracy of localization.

Conclusions

In this paper a WSN based fingerprinting localization method was presented. The RSSI values of the communication links between the previously situated sensors and the mobile sensor were recorded in an indoor environment through the experiment. Using the recorded RSSI values a feed-forward type of neural network was trained. The result of the training is a neural network capable of performing indoor localization. The WkNN algorithm was used as a reference model for the performance verification of the created neural network. The accuracy of the localization between the real and the calculated values was measured with Euclidean distance and demonstrated with the cumulative distribution function. The results have shown that the accuracy of derived neural network also depends on the target position to be determined.

During the testing phase, positions were localized whose RSSI values were not present in the previously recorded database. It has been proven that the accuracy of the applied methods greatly depends on the parameters of the method and the number of the anchors used in the process of localization. In the case of the default structure localization where the RSSI values of all the five anchor motes were processed, the WkNN method proved to be more accurate than the neural network. Since the WkNN method reacted more sensitively to the change of the number of the anchors, when only three anchors were used, the accuracy of the WkNN method was worse than the accuracy of the neural network. This also proves that the position structure of the anchors has a significant effect on the accuracy of localization.

The test results have shown that the number of the anchors and their spatial position have a significant effect on the accuracy of the fingerprinting localization methods discussed. The authors plan on performing more experiments using different anchor structures in their future research.

Acknowledgement

This work was partially supported by the Ministry of Education and Science of the Republic Serbia under the Grant TR-32034, by the Secretary of Science and Technology Development of Vojvodina Province under the Grant 114-451-2434/2011-01 and by the Hungarian Development Agency under the TÁMOP-4.2.2.A-11/1/KONV-2012-0073 program.

References

[1] R. Stoleru, T. He, John A. Stankovic, D.Luebke, “A High-Accuracy, Low-Cost Localization System for Wireless Sensor Networks”, *Proceedings of the 3rd international conference on Embedded networked sensor systems*, pp. 13-26, San Diego, California, 2005

[2] M. Boushaba, A. Hafid, A. Benslimane, “High Accuracy Localization Method Using AoA in Sensor Networks”, *Computer Networks,* Volume 53, Issue 18, 24, pp. 3076-3088, December 2009

[3] M. Yi-Jen, L. Chih-Min, I. J. Rudas, “Wireless Sensor Network (WSN) Control for Indoor Temperature Monitoring ”, *Acta Polytechnica Hungarica*, Vol. 9, No. 6, 2012, pp. 17-28

[4] S. H. Fang and T. N. Lin, “Indoor Location System Based on Discriminant-Adaptive Neural Network in IEEE 802.11 Environments", *IEEE Trans. Neural Networks*, Vol. 19, no. 11, pp. 1973-1978, 2008

[5] L. Gogolak, Sz. Pletl, D. Kukolj, "Indoor Fingerprint Localization in WSN Environment Based on Neural Network", *9th International Symposium on Intelligent Systems and Informatics- SISY2011, IEEE*, pp. 293-296, 8-10 Sept. 2011

[6] J. Simon, G. Martinović, I. Matijevics, “WSN Implementation in the Greenhouse Environment Using Mobile Measuring Station” *International Journal of Electrical and Computer Engineering Systems,* pp. 1-10, Osijek, Croatia, 2010

[7] R. Belbachir, Z. M. Mekkakia and A. Kies, “Towards a New Approach in Available Bandwidth Measures on Mobile Ad Hoc Networks”, *Acta Polytechnica Hungarica*, Vol. 8, No. 4, 2011, pp. 133-148

[8] Sz. Pletl, P. Gál, D. Kukolj, L. Gogolák, “An Optimizing Coverage in Mobile Wireless Sensor Networks” *8th International Symposium on Intelligent Systems and Informatics -SISY 2010,* Subotica, September 2010

[9] L. K. Qabajeh, M. L. M. Kiah and M. M. Qabajeh, “ Secure Unicast Position-based Routing Protocols for Ad-Hoc Networks”, *Acta Polytechnica Hungarica*, Vol. 8, No. 6, 2011, pp. 191-214

[10] D. Kukolj, M. Vuckovic, S Pletl , "Indoor Location Fingerprinting Based on Data Reduction," *Broadband and Wireless Computing, Communication and Applications (BWCCA), 2011 International Conference on* , vol., no., pp. 327-332, 26-28 Oct. 2011

[11] T. Martinetz and K. Schulten. “Topology Representing Networks” *Neural Networks,* Vol. 7, No. 3, pp. 507-522, 1994

[12] D. Kukolj, B. Atlagić, M. Petrov, “Unlabeled Data Clustering Using a Re-organizing Neural Network,” *Cybernetics and Systems*, *An Int. Journal*, Vol. 37, No. 7, pp. 779-790, 2006

[13] G. H Golub and C. F. Van Loan, “Matrix Computations*”*, *Johns Hopkins University Press*, Baltimore, 1989

[14] D. Kukolj, E. Levi, “Identification of Complex Systems Based on Neural and Takagi-Sugeno Fuzzy Model,” *IEEE SMC-part B*, Vol. 34, No. 1, pp. 272-282, February 2004

[15] A. A. Kannan, M. Guoqiang, B. Vucetic, “Simulated Annealing-based Wireless Sensor Network Localization” *Journal of Computers*, Vol. 1, No. 2, pp. 15-22, May 2006

[16] G. Antal,. K. Lamár, “Modern Solutions to Integrated Building Automation Systems” *Proceedings of the International Conference ”Kandó 2002”,* Budapest, Hungary p. 5, 2002

[17] T. Boros, K. Lamár, ”Six-Axis Educational Robot Workcell with Integrated Vision System” *Proceedings of 4th IEEE International Symposium on Logistics and Industrial Informatics ”LINDI 2012”,* Smolenice, Slovakia pp. 239-244, 2012

[18] K. Hechenbichler, K. P. Schliep, “Weighted k-Nearest-Neighbor Techniques and OrdinalClassification” *Discussion Paper 399*, SFB 386, Ludwig-Maximilians University Munich, October 2004

[19] S. H. Fang, C. H. Wang, T. Y. Huang, C. H. Yang, Y. S. Chen, “An Enhanced ZigBee Indoor Positioning System With an Ensemble Approach”, *IEEE Communications Letters,* Vol. 16, No. 4, pp. 564-567, April 2012

VLSI Implementation of High Performance Optimized Architecture for Video Coding Standards

S. Rukmani Devi[1], P. Rangarajan[2], and J. Raja Paul Perinbam[3]

[1)] Electronics and Communication Department, R.M.K. Engineering College, Chennai -601206, India, e-mail: drd.ece@rmkec.ac.in

[2)] Electrical and Electronics Department, R.M.D. Engineering College, Chennai-601206, India, e-mail: hodeee@rmd.ac.in

[3)] Electronics and Communication Department, Karpaga Vinayaka College of Engineering and Technology, Chennai-601206, India, e-mail: jrpp@annauniv.edu

Abstract: This study presents a fast search algorithm and its Very Large Scale Integration (VLSI) design to implement an Enhanced Diamond Search (EDS) of Block-based Motion estimation for video compression systems. The proposed algorithm reduces the number of search points with a slight increase in average Sum-of-Absolute Difference (SAD) per pixel and significant reduction in Peak Signal-to-Noise Ratio (PSNR) than the Full search (FS). The speed improvement ratio is 11.26%. A novel Motion Estimation (ME) algorithm is proposed in this study that exhibits some properties with potential to facilitate optimized VLSI implementation and to achieve high performance for real video sequences. The main characteristic of this proposed architecture is possessing of only five processing elements (PE) that are used to calculate minimum SAD by using efficient comparators and compressors. Simulation results indicate that our proposed architecture employs low power and processes the video data with high speed than existing architectures without significant change in the video quality.

Keywords: Motion Estimation; VLSI; Diamond search; video compression; PSNR; SAD

1 Introduction

Video compression standards is developed by international organizations such as ISO/IEC and ITU-T. ISO/IEC MPEG standard includes MPEG-1, MPEG-2, MPEG-4 Part-2, MPEG-4 Part 10 (AVC), MPEG-7, MPEG-21 and M-JPEG. ITU-T VCEG standard includes H.26x series (H.261, H.263 and H.264) [1]. In attaining the video coding standards, Motion Estimation (ME) plays a vital role to achieve significant compression by exploiting temporal redundancy existing in a video sequence. ME is the most computationally intensive functions of the entire video coding process. Of several algorithms available for block matching motion estimation, full search algorithm identifies the best match in the entire search window by computation of SAD at each location. For a search window of size +/-W pixels, the number of search locations is $(2W+1)^2$[2]. For a search window of 31 x 31 and a block size of 16 x 16, a total of 961 locations will be searched to obtain a best match with a minimum SAD value. This results in significant computational complexity and consumes up to 80% of the computing power of the encoder. To reduce the computational complexity, numerous optimized search algorithm are available as follows: Three Step Search (TSS), New Three Step Search (NTSS), Four Step Search (FSS), Block Based Gradient Descent Search (BBGDS), Diamond Search (DS), Hexagon–Based Search (HEXBS), Adaptive Rood Pattern Search (ARPS), Cross Diamond Search (CDS) [3-12]. The TSS reduces the number of computations by using a coarse-to-fine search strategy. The other algorithms mentioned above reduces the number of computations in relation to TSS by using a center-biased motion vector distribution characteristics. In real world video sequences, the motion of blocks are as stationary or quasi-stationary. If blocks are stationary (about 40%-60%), the corresponding Motion Vectors (MVs) are located at the search center and if the blocks are quasi-stationary (30% -40%), the corresponding MVs are enclosed in a window of size ±2 pixel distance around the center [13-14].

In this paper, we propose an Enhanced Diamond Search (EDS) algorithm for fast block motion estimation algorithm. After analyzing, the characteristics of the MVs, the search pattern of Large Cross Diamond Pattern (LCDP) consists of five search points with pixel distances of ±2 and Small Cross Diamond Pattern (SCDP) consists of four search points with pixel distance of ±1 based on the origin. This proposed algorithm achieves fewer search points and better Peak Signal-to-Noise Ratio (PSNR) than DS, CDS and other fast search algorithms.

The real time processing of video signals requires tremendous computational capabilities that can only be achieved cost effectively by using VLSI. The usefulness of motion estimation algorithms strongly depends on the feasibility and effectiveness of its VLSI implementation. We propose low cost optimized high performance architecture for EDS algorithm used for low bit rate applications.

This paper is organized as follows: Section 2 discusses about proposed algorithm, Section 3 deals with proposed hardware architecture Section 4 describes about synthesis results and conclusion given in the Section 5.

2 Enhanced Diamond Search Algorithm

The EDS algorithm has two cross diamond patterns such as a large cross diamond pattern (LCDP) and small cross diamond pattern (SCDP) as shown in Figure 1(a) and Figure 1(b). It consists of five candidate search points suitable for exploiting the center biased characteristic of motion vector distribution. Figure 2 shows the position of the diamond, with respect to the previous position, for the next search. At subsequent steps, three or two new candidate search points to be evaluated with the maximum overlapping region are required to minimize the number of search points. To obtain the minimum SAD, the final search step with four new candidate search points is evaluated.

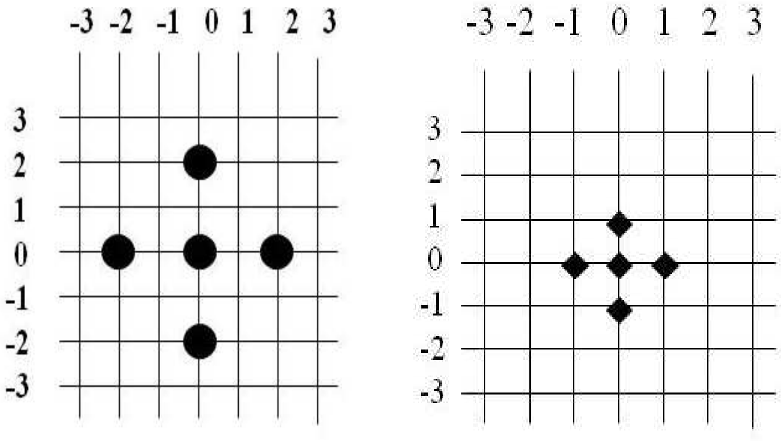


Figure 1 (a)
Large Cross Diamond Pattern (LCDP)

Figure 1(b)
Small Cross Diamond Pattern (SCDP)

Figure 1
The EDS patterns

In order to describe the EDS algorithm, a sum of absolute difference (SAD) obtained by employing the following formula[8]

$$SAD(i,j) = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} |CB(m,n) - RB(m+i, n+j)| \qquad (1)$$

Where N is the block size & CB and RB are the pixel values in the current block and the reference block respectively. PSNR characterizes the motion compensated image created by predicting motion vectors and blocks from the reference frame using the following formula [3]

$$PSNR = 10 \log_{10} \{(2^b - 1)^2/MSE\} \qquad (2)$$

Where b is the number of bits per pixel and MSE is Mean Square Error.

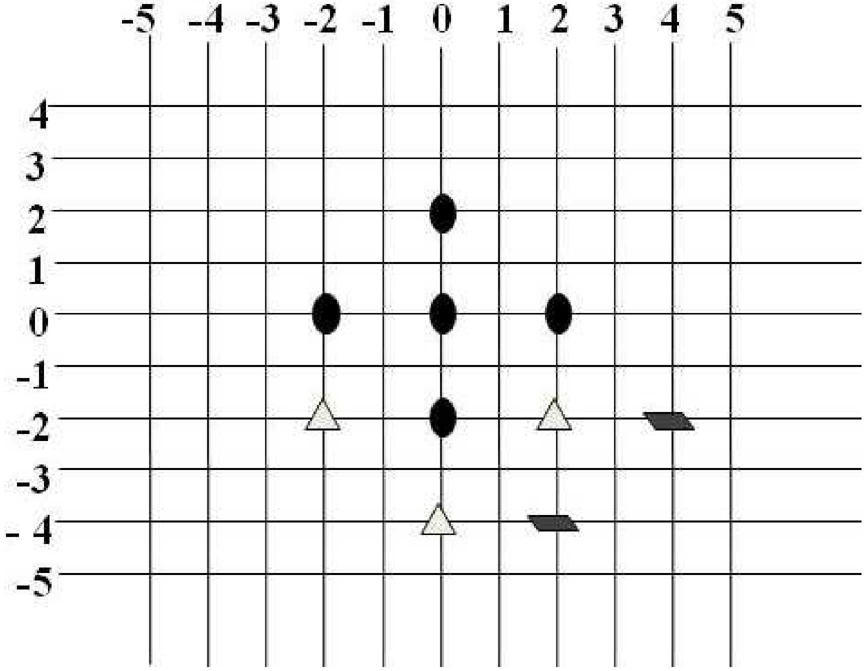


Figure 2
Search path example

The EDS algorithm is summarized as follows:

Step 1. The LCDP is placed at (0, 0) the center of the search window. The SAD is calculated for each of the 5 candidate search points .If the minimum SAD is found to be at the center (a, a) of the LCDP, jump to step 4.

Step 2. If the minimum SAD in the previous search step is located at any one of the vertices i.e. (a±2, a) and (a, a±2), then proceed with step 3.

Step 3. Obtain minimum SAD by repositioning the LCDP considering three and two new candidate search points with a pixel distance of ±2. If the new minimum SAD is still at the center of the newly formed LCDP, proceed to step 4, otherwise continue with step 3. A candidate search point that extends beyond the search window is ignored.

Step 4. A new SCDP is formed, if the minimum SAD is in the center of LCDP. Add four new search points with a pixel distance of ±1 to identify the new minimum SAD. If the minimum SAD is at the center of SCDP, the algorithm stops and corresponding SAD value is considered as the final value of MV. The flowchart of this algorithm is shown in Figure 3.

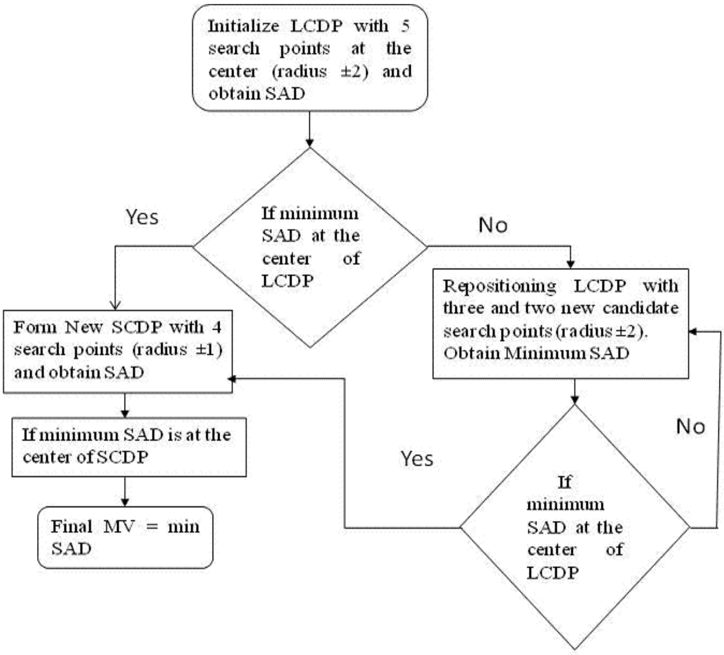

Initialize LCDP with 5 search points at the center (radius ±2) and obtain SAD
If minimum SAD at the center of LCDP
Yes
No
Form New SCDP with 4 search points (radius ±1) and obtain SAD
Repositioning LCDP with three and two new candidate search points (radius ±2). Obtain Minimum SAD
If minimum SAD is at the center of SCDP
Yes
No
If minimum SAD at the center of LCDP
Final MV = min SAD

Figure 3
Flowchart of the EDS algorithm

2.1 Software Analysis

A software analysis was performed to evaluate the average number of search points per block, average SAD per pixel and PSNR using MATLAB and results are compared for two different video sequences as given in the Table 1. The search area employed was 32 x 32 pixels with block size of 16 x 16 pixels. The algorithm was applied to the first 30 frames of CIF video sequences with a resolution of 352 x 288 pixels.

Table 1
Comparison of DS and proposed algorithm for two video sequences

Sequences	DS			**Proposed**		
	Average No. of search points / Block	Average SAD / Pixel	Peak Signal-to-Noise Ratio (PSNR)	**Average No. of search points / Block**	**Average SAD / Pixel**	**Peak Signal-to-Noise Ratio (PSNR)**
Foreman	13.99	5.920	33.591	**10.30**	**5.996**	**33.615**
Mobile	17.51	11.734	32.582	**11.62**	**11.957**	**32.699**

3 Proposed Hardware Architecture

The proposed architecture consists of current block memory, reference block memory, data-fetch unit, data-fetch initializer, Processing Element (PE) array, comparator, PE array Enabler and Timing and control unit as shown in Figure 4. To start Motion Estimation (ME) process, it is necessary to fill the reference block memory and the current memory block. The reference block memory is 32 x 32 bytes memory, which can store all data to perform SAD operations and the final refinement. The current block memory is 16 x 16 bytes and it contains the block being processed. The data-fetch initializer initializes the data-fetch unit to fetch reference pixels and current pixels from the memory, row and column address of the block at the center of the reference frame and type of fetching i.e. pixels of horizontal or vertical direction .It also calculates starting address of the values to be fetched, number of search points to be fetched at a time and number of cycles needed to fetch data input. After initializing, the data-fetch unit obtains the pixel values from the current and reference block. The pixel values are applied to the PE array. A PE array with 5 processing elements is used to calculate SAD between the current block and reference block. Totally five SAD values are generated by the PE array and these are compared by the comparator to evaluate minimum SAD or best motion vector and position of the motion vector. The process is repeated until the comparator finds the lowest SAD in the center of the EDS. Low power can be achieved by using the PE array enabler. This block is used to enable appropriate processing element based on the search path. The timing and control unit is a finite state machine, which is used to control all the blocks presented in the proposed architecture.

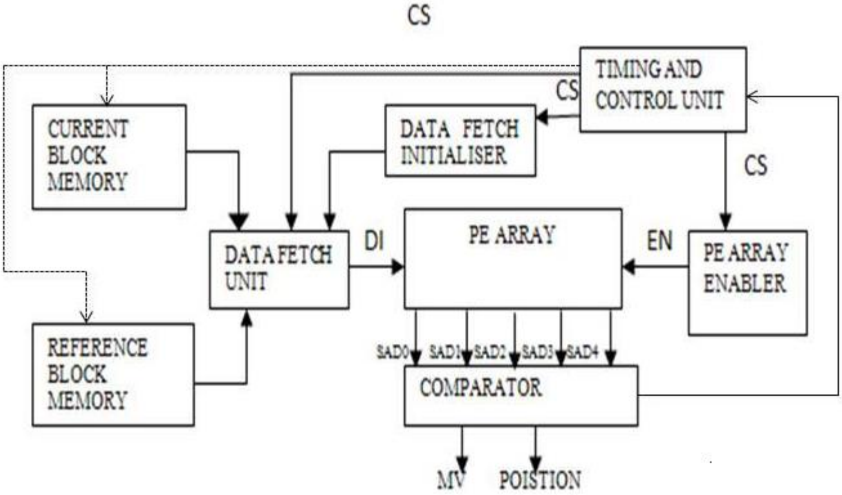


Figure 4

Proposed Optimized Architecture of the EDS algorithm

3.1 Processing Element

The PE array was designed with five processing elements and four stage pipeline for better performance. It consists of an absolute difference unit, adder array and accumulator as shown in Figure 5. The internal structure of the PE is composed of a large number of addition operations to calculate the SAD. These operations can be performed effectively by using 5-2 and 3-2 adder compressors [15-19] as shown Figure 8(a).

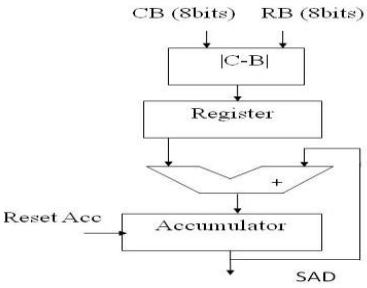


Figure 5
Processing Element

The absolute difference unit is used to evaluate the absolute differences between two pixels. It consists of eight 1-bit comparators (1-bit CMP), seven 2-bit comparators (2-bit CMP) and two XOR arrays [15-19] as shown in Figure 6. The 1-bit CMP is formed by a XOR gate and two AND gates. It produces three output signals: they are less than (LE), Equal to (NE) and greater than (LG), as shown in Figure 7 (a). The 2-bit comparator is formed by using two 1-bit CMP as shown in Figure 7 (b).

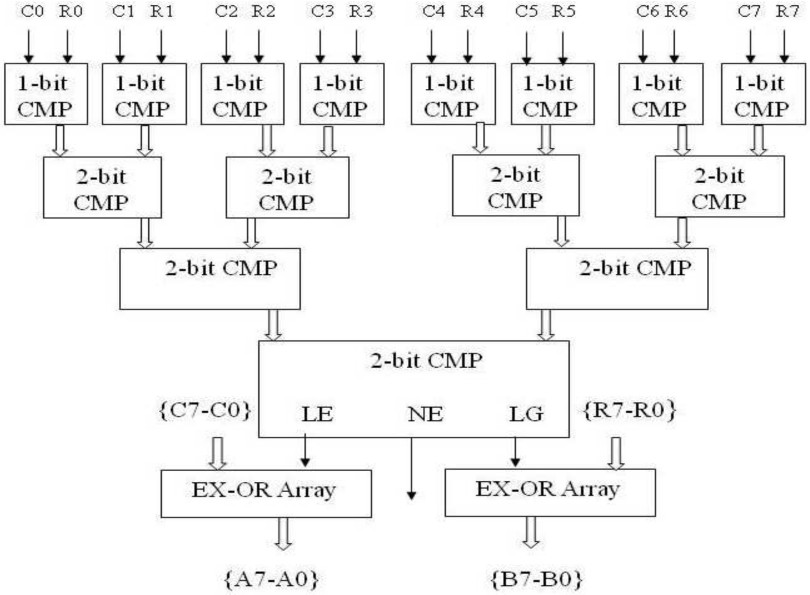
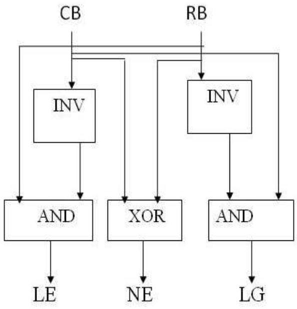


Figure 6

Absolute Difference Unit (Proposed by LiYufei et.al)



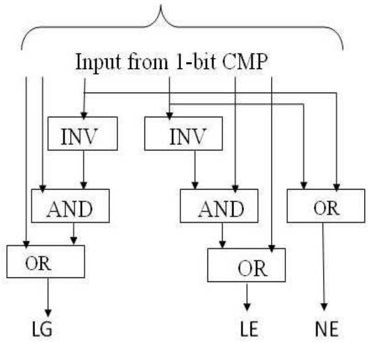


Figure 7 (a)

1-bit Comparator

Figure 7 (b)

2-bit Comparator

Processing element (PE) performs a large number of additions that can be optimized by using 5-2 and 3-2 compressor for an 8-bit number as shown in Figure 8. The internal structure of 5-2 and 3-2 compressors are shown in Figure 9(a) and Figure 9(b).

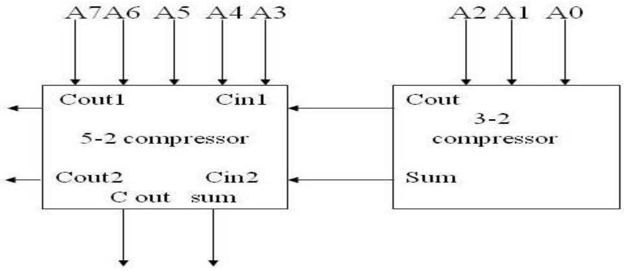
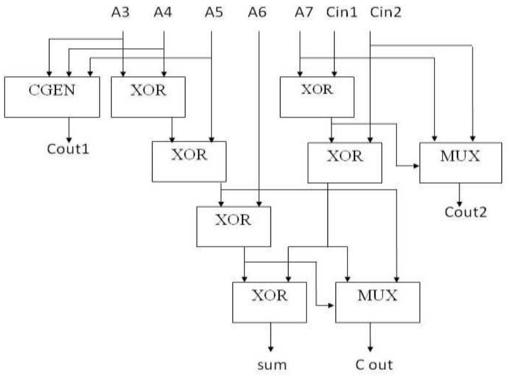


Figure 8

5-2 Compressor and 3-2 Compressor



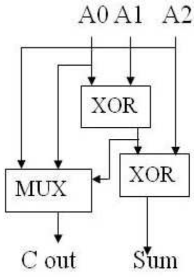


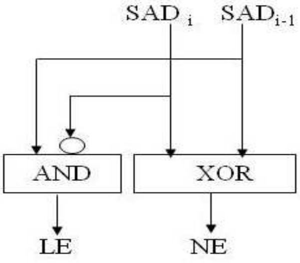
Figure 9 (a)

Internal structure of 5-2 compressor

Figure 9 (b)

Internal structure of 3-2 compressor

Depending on the search path, the PE generates SAD values. These values are compared by using comparator (as shown in figure 4) and minimum SAD is obtained. For optimization purposes, the comparator is constructed using 1-bit less comparator (1-bit LCMP) and 2-bit less comparator (2-bit LCMP) as shown in Figure 10(a) and Figure 10(b).



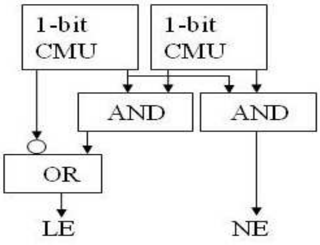


Figure 10 (a)

1-bit LCMP

Figure10(b)

2-bit LCMP

3.2 Timing and Control Unit

The timing and control unit generates a control signal (CS) to control all the blocks present in the proposed architecture (refer Figure 4). The control unit is a Finite State Machine (FSM) that controls the operation of each unit based on the search path. The flow remains to be linear owing to simpler control of the architecture. Table 2 explains the operation of the control unit.

Table 2
Timing and Control signal

State	State name	Description
S0	CLEAR	Clears PE, ACC, DF unit, DF initialize, PE array enabler & waits for a clear signal to move to the next state.
S1	EXTERNAL INPUT STATE (DI)	Receives and stores the external inputs
S2	HORIZONTAL PHASE STATE	Selects search type and starts horizontal computation
S3	SAD COMPUTATION STATE	Starts the sad computation by making enable (EN) =1
S4	COMPARATOR INITIALISE STATE	Comparison of the SAD values for horizontal pixels to find minimum SAD
S5	HORIZONTAL COMPLETE STATE	Clears PE, ACC, and DF unit PE array enabler.
S6	VERTICAL PHASE STATE	Selects search type and starts vertical computation
S7	SAD COMPUTATION STATE	Starts the sad computation for vertical by making enable (EN) =1
S8	COMPARATOR INITIALISE STATE	Comparison of the SAD values for vertical pixels to find minimum SAD
S9	FINISH	If minimum SAD is obtained go to the next block

4 Synthesis Results

The proposed architecture was implemented in Verilog HDL using Xilinx 12.2 in Spartan 6 SP605 FPGA LX45T kit and output was analyzed using Chip Scope Pro. Table 3 presents the synthesis results of this proposed algorithm for two different devices. The device hardware utilization is small, since only 3214 look-up-tables (FPGA LUTs) are used. The performance results presented in Table 3 take into account CIF (352 x 288 pixels) videos. This hardware consumes 1711 slices, which is 27% of all the slices of an xc6slx45Tfgg484-3 with maximum frequency of 397.844 MHz and power consumption of 120 mw with minimum latency. Table 4 gives the comparison of the proposed architecture with other architectures.

Table 3
Synthesis Results

Parameters	**EDS Results**	
	(xc6slx45Tfgg484-3)	(2v40cs144-5)
No. of Slices	1711 (27%)	87(33%)
No. of Flip flops	384 (3%)	30(5%)
No. of 4 I/P LUTs	3214 (26%)	156(30%)
Minimum Period	2.514ns	4.227ns
Maximum frequency (MHz)	397.84	236.574

Table 4 compares various parameters between the proposed and existing diamond search algorithms.

Table 4
Performance Comparison of Proposed with other architectures

Parameters	QSDS [15]	SDS[20]	**Proposed**
No. of Slices	2007	1968	**1711**
No. of Flip flops	2086	2020	**384**
No. of 4 I/P LUTs	3610	3541	**3214**
Minimum Period	4.688 ns	-	**2.514 ns**
Maximum frequency (MHz)	213.3	185.7	**397.84**
Number of PE	9	9	**5**
Search range	64 x 64	64 x 64	**32 x 32**

Conclusion

The enhanced diamond search algorithm displayed better performances as compared to full search and other fast search algorithms. This algorithm performs a less number of search points and 0.024 dB increase in PSNR for medium motion sequence (Foreman) and 0.177 dB increase in object translation sequences (Mobile) with 11.26% speed improvement ratio than other algorithms. In this study, a VLSI architecture was developed for this algorithm, which uses only five processing elements to maintain 30 f/s with an operating frequency of 397 MHz. Totally, 32 clock cycles per block needed to obtain a SAD value. To improve the performance of this architecture, optimized PE was incorporated. This algorithm is developed for fixed block size motion estimation and was implemented successfully in FPGA device. This work could be extended to variable block size motion estimation for an H.264 AVC standard.

References

[1] Bhaskaran and K. Konstantinides, Image and Video Compression Standards Algorithms and Architectures, Kluwer Academic, Boston, Mass, USA, 1999

[2] S.-Y. Huang. Chuan-Yu Cho ; Jia-Shung Wang "Adaptive Fast Block-Matching Algorithm of Switching Search Patterns for Sequences With Wide-Range Motion Content", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 15, No. 11, pp. 1383-1384, November 2005

[3] Shiping Zhu, Jun Tian, Xiaodong Shen, Kamel Belloulata," A New Cross-Diamond Search Algorithm for Fast Block Motion Estimation", pp. 1581-1584, ICIP-2009

[4] www.cwaip.nus.edu.sg

[5] Sanchez, Gustavo, Felipe Sampaio, Marcelo Porto, Sergio Bampi, and Luciano Agostini. "DMPDS: A Fast Motion Estimation Algorithm Targeting High Resolution Videos and Its FPGA Implementation", International Journal of Reconfigurable Computing, pp. 1-12, October 2012

[6] "IEEE-ICDCS conference proceeding", 2012 International Conference on Devices Circuits and Systems (ICDCS), 03/2012

[7] www.ee.cityu.edu.hk

[8] Reeba Korah, Sankaralingam and M. J. R. P. Perinbam, "Motion Estimation with Candidate Block and Pixel Subsampling Algorithm", IEEE International Workshop on Imaging Systems and Techniques, pp. 130-133, May 2005

[9] Porto. M, Agostini, L. , Bampi, S. and Susin, A. "A High Throughput and Low Cost Diamond Search Architecture for HDTV Motion Estimation", 2008 IEEE International Conference on Multimedia and Expo, pp 1033-1036, June 2008

[10] www.ee.vt.edu

[11] Su-Bong Hong, Hyoseok Lee, Geun-Young Chun, Hyunki Baik and Myong-Soon Park, "FCBHS: a Fast Center-biased Hybrid Search Algorithm for Fast Block Motion Estimation", Proceedings International Conference on Information Technology Coding and Computing ITCC, pp. 254-259, Feb 2002

[12] Yong Ding, Xiao-Lang Yan, "Parallel Architecture of Motion Estimation for Video Format Conversion with Center-biased Diamond Search", International Conference on Information Engineering and Computer Science, pp. 1-4, Dec 2009

[13] Huang, Hui-Yu, and Shih-Hsu Chang. "Block Motion Estimation Based on Search Pattern and Predictor", IEEE Symposium on Computational Intelligence For Multimedia Signal and Vision Processing, pp. 47-51, April 2011

[14] Mahid Asefi, Mohamed-yahia Dabbagh. "Adaptive Video Motion Estimation Algorithm via Estimation of Motion Length Distribution and

Bayesian Classification", IEEE International Symposium on Signal Processing and Information Technology, pp. 807-810, August 2006

[15] Marcelo Porto, André Silva, Sergio Almeida Eduardo DA Costa, Sergio Bampi," Motion Estimation Architecture Using Efficient Adder-Compressors for HDTV Video Coding", Journal Integrated Circuits and Systems, Vol. 5, No. 1, pp. 78-88, 2010

[16] LiYufei , Feng Xiubo and Wang Qin," A High-Performance Low Cost SAD Architecture for Video Coding', IEEE Transactions on Consumer Electronics, pp. 535-541, Vol. 53, No. 2, May 2007

[17] Jarno, Vanne,Eero Aho, Timo D. Hämäläinen, and Kimmo Kuusilinna, "A High-Performance Sum of Absolute Difference Implementation Motion Estimation, IEEE Transactions on Circuits and Systems for Video Technology, pp. 876-883, Vol. 16, No. 7, 2006

[18] Chip-Hong Chang, Jiangmin GU, and Mingyan Zhang," Ultra Low-Voltage Low-Power CMOS 4- 2 and 5-2 Compressors for Fast Arithmetic Circuits", IEEE Transactions on Circuits and Systems—I: Regular Papers, pp. 1985-1997, Vol. 51, No. 10, October 2004

[19] Meihua Gu, Ningmei Yu, Lei Zhu, Wenhua Jia," High Throughput and Cost Efficient VLSI Architecture of Integer Motion Estimation For H.264/AVC", Journal of Computational Information Systems, pp. 1310-1318, Vol. 7, No. 4, April 2011

[20] Marcelo Porto, Luciano Agostini, Sergio Bampi, AltamiroSusin," A High throughput And Low Cost Diamond Search Architecture For HDTV Motion Estimation", pp. 1033-1036, ICME 2008